

# Asynchronous Optimization Over Heterogeneous Networks via Consensus ADMM

Sandeep Kumar, *Student member, IEEE*, Rahul Jain, and Ketan Rajawat, *Member, IEEE*

**Abstract**—This paper considers the distributed optimization of a sum of locally observable, non-convex functions. The optimization is performed over a multi-agent networked system, and each local function depends only on a subset of the variables. An asynchronous and distributed alternating directions method of multipliers (ADMM) method that allows the nodes to defer or skip the computation and transmission of updates is proposed in the paper. The proposed algorithm utilizes different approximations in the update step, resulting in proximal and majorized ADMM variants. Both variants are shown to converge to a local minimum, under certain regularity conditions. The proposed asynchronous algorithms are also applied to the problem of cooperative localization in wireless ad hoc networks, where it is shown to outperform the other state-of-the-art localization algorithms.

**Index Terms**—non-convex problems, asynchronous algorithms, distributed optimization, majorization, ADMM

## I. INTRODUCTION

Multi-agent networked systems arise in a number of engineering disciplines such as tactical ad hoc networks [1], environmental monitoring networks [2], multi-robot control and tracking [3], [4], internet-scale monitoring [5], [6], and large-scale learning [7]. The estimation, resource allocation, and network control tasks required in these applications are often formulated as distributed optimization problems, where each node is associated with a local cost function, determined from set of local and possibly private measurements [8], [9]. The nodes must nevertheless cooperate in order to minimize the network objective function, which is the sum of local costs. The algorithm design becomes challenging in the absence of a centralized controller or a fusion center, where nodal interactions are limited to their neighborhoods, and global network state information is largely unavailable [8], [10]. In high-dimensional problems, even local message passing may be prohibitive, since each node may only be interested in a subset of the optimization variables.

In general, distributed optimization algorithms are designed either in the primal [9], [11], [12], [13, Chap. 10] or dual domain [8], [14]. A popular dual approach is the distributed alternating direction method of multipliers (ADMM), where the nodal variables are decoupled through the introduction of the consensus constraints, and the updates are carried out in the dual domain [8], [15]. High-dimensional problems are handled through the so-called *general-form consensus* formulation, where local updates depend only on a subset of

optimization variables [8, Chap. 7]. The ADMM algorithm is also applicable to a class of non-convex problems, where it has been shown to converge to a local minimum [16], [17].

Practical networks are also heterogeneous with respect to their processing powers, energy availability, and communication capabilities, giving rise to asynchrony [18], [19]. Indeed, a key feature required of the distributed algorithms is their tolerance to processing and communication delays arising due to slow or energy-starved nodes [15], [20]–[27]. In general, distributed optimization algorithms such as ADMM must be appropriately modified to allow updates to be skipped or delayed, and the convergence of the asynchronous variants is neither obvious, nor guaranteed [16], [18, Chap. 6]. Asynchronous or randomized variants of the distributed ADMM are well-known for the case when the cost functions are convex; see e.g., [15], [25], [28] and references therein. Likewise, if a master node or a fusion center is available, the asynchronous distributed ADMM variant proposed in [16], is applicable to non-convex cost functions and has been shown to converge to a local optimum.

This work considers the non-convex general-form consensus optimization problem arising in a multi-agent networked system. The first contribution is the development of an asynchronous and distributed ADMM framework that runs without a fusion center. Two variants, namely, the proximal and the majorized ADMM, are proposed, each handling the non-convex objective function in a different way. While the ADMM iterations at all nodes still occur according to a common schedule, both algorithms allow the nodes to, at times, skip the computationally intensive steps and/or the transmission of updates. As the second contribution, it is shown that both variants converge to a local minimum, under certain regularity conditions. The convergence analysis reveals that with appropriately chosen parameters, the algorithms can tolerate any bounded level of asynchrony. Finally, the third contribution is the application of the proposed ADMM algorithm to the problem of cooperative localization for distributed networks [3], [4]. Detailed simulations and comparisons with existing distributed and asynchronous localization algorithms are carried out, establishing the superior performance of the ADMM algorithm.

This paper is organized as follows. The problem formulation and related examples are provided in Sec. II. The proposed proximal asynchronous ADMM and the associated convergence results are presented in Sec. III, while the proposed majorized asynchronous ADMM is detailed in Sec. IV. Finally, the simulation results for the proposed algorithm applied to the cooperative localization problem are provided in Sec. V and

Sandeep Kumar and Ketan Rajawat are with the Dept. of Electrical Engineering, IIT Kanpur, Kanpur, UP-208016, India; emails: {sandkr,ketan}@iitk.ac.in. Rahul Jain is with Numerify.com, Bangalore Karnataka-560025, India; email rahuljain2104@gmail.com

sec. VI concludes the paper.

## II. PROBLEM FORMULATION

This section details the partially separable non-convex problem formulation considered here, and motivates the need for a distributed optimization algorithm via two examples. Before describing the problem at hand, some background regarding the general multi-agent optimization problem is first presented.

### A. Background

Consider a network represented by the undirected graph  $\mathcal{G} = (\mathcal{K}, \mathcal{E})$ , where  $\mathcal{K} := \{1, 2, \dots, K\}$  denotes the set of agents or nodes and  $\mathcal{E}$ , the set of edges that represent communication links. A node  $k \in \mathcal{K}$  may only communicate with its neighbors  $\mathcal{N}'_k := \{j | (j, k) \in \mathcal{E}\}$ . Consider first the general multi-agent optimization problem where the nodes want to cooperatively solve the following optimization problem:

$$\min_{\mathbf{x}} f(\mathbf{x}) := \sum_{k=1}^K g_k(\mathbf{x}) + h(\mathbf{x}) \quad \text{s. t. } \mathbf{x} \in \mathcal{X} \quad (1)$$

where  $\mathbf{x} \in \mathbb{R}^{N \times 1}$ ,  $g_k : \mathbb{R}^N \rightarrow \mathbb{R}$  for  $k = 1, \dots, K$  are differentiable, possibly non-convex functions, and  $h : \mathbb{R}^N \rightarrow \mathbb{R}$  is a convex but not necessarily differentiable function. The set  $\mathcal{X}$  is closed, convex, and compact. Within the distributed setting considered here, the function  $g_k(\mathbf{x})$  is local to node  $k$ , and the network has no central coordinator or fusion center.

In general, non-convex problems such as (1) are solved in a distributed manner using the first order gradient or subgradient descent, dual methods such as ADMM, convex relaxation (such as semidefinite relaxation), successive convex optimization, or leveraging the problem structure; see [29] and references therein. These approaches result in algorithms that are parallelizable to various extents, with different computational and message passing requirements. Of particular interest here is the high-dimensional regime, where large  $N$  prohibits nodes from operating over and exchanging the full vector  $\mathbf{x}$ . To this end, the next section considers a partially separable form of (1) which is amenable to a distributed optimization algorithm.

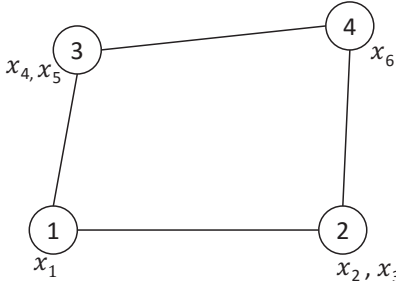


Fig. 1: An example of a four-node network.

### B. Partially separable form

Consider a special case of (1) where the optimization variables  $\{x_n\}_{n=1}^N$  are also partitioned among nodes, and each variable is of interest to exactly one node. To this end, let  $\{S_k\}_{k=1}^K$  denote disjoint subsets such that the variables

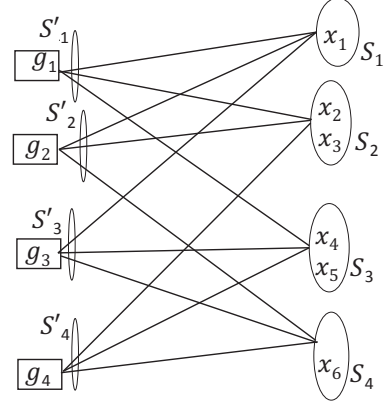


Fig. 2: Factor graph representation for the objective function of (2).

$\{x_n | x_n \in S_k\}$  are local to node  $k$ . Further, the component function  $g_k(\cdot)$  at node  $k$  depends only on variables that are local either to node  $k$  or its neighbors. The overall problem considered here takes the following form:

$$P = \min_{\mathbf{x}} \sum_{k=1}^K g_k(\{x_n\}_{n \in S'_k}) + h_k(\{x_n\}_{n \in S_k}) \quad (2)$$

$$\text{s. t. } \{x_n\}_{n \in S_k} \in \mathcal{X}_k \quad k = 1, 2, \dots, K.$$

where the set  $S'_k := \bigcup_{j \in \{k\} \cup \mathcal{N}'_k} S_j$ . Observe here that for each node  $k$ , the function  $h_k$  and the constraint set  $\mathcal{X}_k$  depend only on the variables in  $S_k$ . For this partially separable form, it is now possible to express the objective function as a bipartite factor graph, with check nodes representing the summands  $g_k$ , and the variable nodes representing the sets  $S_k$ . For the four-node example network shown in Fig. 1, the factor graph is shown in Fig. 2. From the perspective of algorithm design, the dependence structure imposed by (2) can be exploited to eliminate message passing between non-neighboring nodes. The partially separable form considered here occurs commonly in the context of distributed estimation, where the parameter of interest is a collection of node-specific quantities such as temperatures, node locations, harmful algal blooms, pH, and temperature [30], [31]. Cooperation between nodes is still required here, since parameters at neighboring nodes are often coupled or correlated. The subsequent examples detail two specific applications where the partially separable form problem structure in (2) arises.

### C. Examples

**Example 1.** Consider a sensor network with nodes  $\mathcal{K} := \{1, 2, \dots, K\}$  observing an environmental phenomenon. The environmental state at the location of sensor  $k \in \mathcal{K}$  is denoted by the random variable  $\theta_k \in \mathbb{R}$ , and is collected in the vector  $\boldsymbol{\theta} \in \mathbb{R}^K$ . The measurements at sensor  $k$  are in form of a vector  $\mathbf{y}_k \in \mathbb{R}^{N_k}$  and depend on the local state  $\theta_k$  via the pdf  $p(\mathbf{y}_k | \theta_k)$ . Given the overall measurement vector  $\mathbf{y}$ , that collects the nodal measurements  $\{\mathbf{y}_k\}_{k=1}^K$ , the goal at each node is to calculate the *maximum a posteriori* (MAP) estimate of  $\theta_k$ , given by  $\hat{\theta}_k := \arg \max_{\theta} p(\theta | \mathbf{y}) = \arg \max_{\theta} \ln p(\boldsymbol{\theta}) + \sum_{k=1}^K \ln p(\mathbf{y}_k | \boldsymbol{\theta})$ , where  $p(\boldsymbol{\theta})$  captures the available prior information about  $\boldsymbol{\theta}$ . Of particular

interest here are problems where the prior pdf has a factored representation, i.e.,  $p(\boldsymbol{\theta}) = \prod_{k=1}^N p_k(\boldsymbol{\theta}_{C_k})$ , where  $C_k \subset \mathcal{K}$ , and  $\boldsymbol{\theta}_{C_k}$  stacks the variables  $\{\theta_n\}_{n \in C_k}$ . Such factored pdfs arise naturally within the context of distributed sensing of Markov random fields [32], where  $C_k = \mathcal{N}_k$ . The Markov property implies that  $\theta_k$  is correlated with  $\{\theta_n\}_{n \in \mathcal{N}_k}$ , but is conditionally independent of  $\{\theta_n\}_{n \notin \mathcal{N}_k}$ , given  $\{\theta_n\}_{n \in \mathcal{N}_k}$ ; see e.g [33] and references therein. When the pdf can be factored, the MAP estimation problem may be written as

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} \sum_{k=1}^K \ln p_{N_k}(\boldsymbol{\theta}_{N_k}) + \sum_{k=1}^K \ln p(\mathbf{y}_k | \boldsymbol{\theta}_k). \quad (3)$$

It can be observed that (3) is a special case of (2) with  $g_k(\cdot) = \ln p_k(\cdot)$  and  $h_k(\cdot) = \ln p(\mathbf{y}_k | \cdot)$ .

**Example 2.** This example details the problem of cooperative localization in wireless networks via multidimensional scaling. Distributed localization is necessary in networks where manual or GPS-assisted positioning is not feasible, and not all nodes can communicate with the fusion center or base stations. Within the cooperative localization framework considered here, the nodes utilize pairwise range measurements and the known location of a few anchor nodes to estimate their locations. Range measurements are often made using techniques such as received signal strength or time-of-arrival [34]. Specifically, consider a network with  $N$  nodes, distributed in a  $p$ -dimensional space with  $p = 2$  or  $3$ . The location of node  $k$  is denoted by  $\mathbf{x}_k \in \mathbb{R}^p$ , and collected into the matrix  $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N]$ . The measured distance between a pair of neighboring nodes  $k$  and  $j$  is denoted by  $\delta_{kj}$ , and is generally a noisy version of  $d_{kj}(\mathbf{x}_k, \mathbf{x}_j) := \|\mathbf{x}_k - \mathbf{x}_j\|$ . The MDS formulation finds the overall configuration  $\mathbf{X}$  by solving the following sum of squares problem [3], [4]

$$\hat{\mathbf{X}} = \arg \min_{\mathbf{X} \in \mathbb{R}^{n \times p}} \sum_{k=1}^N g_k(\{\mathbf{x}_j\}_{j \in \mathcal{N}_k}) \quad (4)$$

where,  $g_k(\{\mathbf{x}_j\}_{j \in \mathcal{N}_k}) = \sum_{j \in \mathcal{N}_k'} w_{kj} (\delta_{kj} - d_{kj}(\mathbf{x}_k, \mathbf{x}_j))^2$ . When the location  $\mathbf{x}_k^p$  of node  $k$  is known a priori, a regularization term of the form  $r_i \|\mathbf{x}_k - \mathbf{x}_k^p\|$  may also be added to  $g_k(\cdot)$ . Since (4) is non-convex, it is often solved via majorization [4] or via SDP relaxation [3]. A distributed and incremental algorithm for solving (4) via majorization was first detailed in [4]. A distributed and asynchronous algorithm using SDP relaxation was described in [3]. The subsequent sections describe the proximal and majorized ADMM algorithms that are also applicable to (4).

### III. DISTRIBUTED ASYNCHRONOUS ADMM

This section details the proposed distributed asynchronous algorithm for solving (2) via ADMM, and provides the relevant convergence results. To begin with, the next subsection describes a distributed synchronous implementation, which is motivated from the so called general-form consensus algorithm for solving convex optimization problems [8]. The synchronous version serves as a starting point for the asynchronous algorithm described in Sec. III-B.

#### A. Distributed Synchronous ADMM

For ease of exposition, it is assumed henceforth that the sets  $\mathcal{S}_k$  are singletons, i.e.,  $\mathcal{S}_k = \{k\}$ . Note however that the algorithms and the corresponding convergence results are applicable to the general case as well. For instance, it is possible to have  $|\mathcal{S}_k| = n_k$  with  $\sum_k n_k = N$  and  $g_k : \mathbb{R}^{N_k} \rightarrow \mathbb{R}$  where  $N_k = n_k + \sum_{j \in \mathcal{N}_k'} n_j$ . Indeed, for the localization example described in Sec. II-C,  $n_k = 2$  for all  $k$ . The expanded neighborhood of node  $k$  is defined as  $\mathcal{N}_k := \{k\} \cup \mathcal{N}_k'$  and includes  $k$  itself. Towards developing a distributed algorithm for solving (2), introduce copies  $x_{kj}$  of the variable  $x_k$  corresponding to all nodes  $j \in \mathcal{N}_k$  for all  $k \in \mathcal{K}$ . To make sure that the copies of the a variable agree across nodes, also introduce the so called consensus variable  $z_k$  for each neighborhood  $\mathcal{N}_k$ . The introduction of these extra variables amounts to reformulating (2) as

$$\min_{\{\mathbf{x}_k\}, \mathbf{z}} \sum_{k=1}^K g_k(\{x_{kj}\}_{j \in \mathcal{N}_k}) + h_k(z_k) \quad (5)$$

$$\text{s. t. } x_{kj} = z_j, \quad j \in \mathcal{N}_k, \quad k = 1, \dots, K \quad (6)$$

$$z_k \in \mathcal{X}_k, \quad k = 1, \dots, K$$

where  $\mathbf{x}_k$  collects the variables required at node  $k$ , i.e.,  $[\mathbf{x}_k]_j = x_{kj}$  for  $j \in \mathcal{N}_k$ , and zero otherwise.

The idea of introducing consensus variables, in order to make the updates separable in the optimization variables, is well known [10, Chap. 5]. It is now possible to apply the ADMM method by associating dual variables  $y_{kj}$  for each constraint in (6) and writing the augmented Lagrangian as

$$L(\{\mathbf{x}_k\}, \mathbf{z}, \{\mathbf{y}_k\}) = \sum_{k=1}^K \left( g_k(\mathbf{x}_k) + h_k(z_k) + \sum_{j \in \mathcal{N}_k} \langle y_{kj}, x_{kj} - z_j \rangle + \sum_{j \in \mathcal{N}_k} \frac{\rho_k}{2} \|x_{kj} - z_j\|^2 \right) \quad (7)$$

where  $\rho_k > 0$  is a positive penalty parameter and  $\mathbf{z}$  collects  $\{z_k\}_{k \in \mathcal{K}}$ . Here, although  $\{x_{kj}\}$ ,  $\{z_j\}$ , and  $\{y_{kj}\}$  are all scalar variables, the notations for generalized inner product and norm are retained in order to avoid manipulations that are specific to the scalar case. On the left-hand side, for each  $k$ , the dual vector  $\mathbf{y}_k \in \mathbb{R}^N$  is such that  $[\mathbf{y}_k]_j = y_{kj}$  if  $j \in \mathcal{N}_k$ , and zero otherwise.

From (7), it is clear that  $L$  is separable in  $\{\mathbf{x}_k\}$ . The Lagrangian is also separable in  $\{z_j\}$ , since it holds that

$$L(\{\mathbf{x}_k\}, \mathbf{z}, \{\mathbf{y}_k\}) = \sum_{j=1}^K \left( g_j(\mathbf{x}_j) + \sum_{k \in \mathcal{N}_j} \langle y_{kj}, x_{kj} - z_j \rangle + \sum_{k \in \mathcal{N}_j} \frac{\rho_k}{2} (x_{kj} - z_j)^2 + h_j(z_j) \right). \quad (8)$$

Together, (7) and (8) allow the ADMM updates to be carried out in a distributed fashion. In particular, starting with arbitrary  $\{\mathbf{x}_k^1\}$  and  $\{\mathbf{y}_k^1\}$ , the update for  $\{z_j^{t+1}\}$  are evaluated as

$$z_j^{t+1} = \arg \min_{z_j \in \mathcal{X}_j} h_j(z_j) + \sum_{k \in \mathcal{N}_j} \langle y_{kj}^t, x_{kj}^t - z_j \rangle + \sum_{k \in \mathcal{N}_j} \frac{\rho_k}{2} \|x_{kj}^t - z_j\|^2$$

$$= \text{prox}_j \left( \frac{\sum_{k \in \mathcal{N}_j} (\rho_k x_{kj}^t + y_{kj}^t)}{\sum_{k \in \mathcal{N}_j} \rho_k} \right) \quad (9)$$

where the proximal point function  $\text{prox}_j(\cdot)$  is defined as

$$\text{prox}_j(x) := \arg \min_{u \in \mathcal{X}_j} h(u) + \frac{1}{2} \|x - u\|^2. \quad (10)$$

Similarly updates for  $\mathbf{x}_k^{t+1}$  can be obtained by minimizing (7) with respect to  $\mathbf{x}_k$ , which yields

$$\begin{aligned} \mathbf{x}_k^{t+1} = \arg \min_{\mathbf{x}_k} g_k(\mathbf{x}_k) + \sum_{j \in \mathcal{N}_k} \langle y_{kj}^t, x_{kj} - z_j^{t+1} \rangle \\ + \sum_{j \in \mathcal{N}_k} \frac{\rho_k}{2} \|x_{kj} - z_j^{t+1}\|^2 \end{aligned} \quad (11)$$

where the optimization is with respect to  $\{x_{kj}\}_{j \in \mathcal{N}_k}$ . Observe that since the component functions  $g_k(\cdot)$  are not necessarily convex, the update in (11) is difficult to carry out. As suggested in [16], the update  $\mathbf{x}_k^{t+1}$  can however be calculated approximately as follows

$$\begin{aligned} \mathbf{x}_k^{t+1} \approx \arg \min_{\mathbf{x}_k} g_k(\mathbf{z}_k^{t+1}) + \langle \nabla g_k(\mathbf{z}_k^{t+1}), \mathbf{x}_k - \mathbf{z}_k^{t+1} \rangle \\ + \sum_{j \in \mathcal{N}_k} \langle y_{kj}^t, x_{kj} - z_j^{t+1} \rangle + \sum_{j \in \mathcal{N}_k} \frac{\rho_k}{2} \|x_{kj} - z_j^{t+1}\|^2 \end{aligned} \quad (12)$$

where the vector  $[\mathbf{z}_k]_j := z_j$  for all  $j \in \mathcal{N}_k$  and zero otherwise. Since nodal functions  $g_k(\cdot)$  depend only on  $\{x_n\}_{n \in \mathcal{N}_k}$ , the gradient vector is defined as

$$[\nabla g_k(\mathbf{z}_k^{t+1})]_j := \begin{cases} \frac{\partial}{\partial x_{kj}} g_k(\mathbf{x}_k) \Big|_{\mathbf{x}_k = \mathbf{z}_k} & j \in \mathcal{N}_k \\ 0 & j \notin \mathcal{N}_k. \end{cases} \quad (13)$$

The approximate update of  $x_{kj}^{t+1}$  thus becomes

$$x_{kj}^{t+1} = \begin{cases} z_j^{t+1} - \frac{1}{\rho_k} \left( [\nabla g_k(\mathbf{z}_k^{t+1})]_j + y_{kj}^t \right) & j \in \mathcal{N}_k \\ 0 & j \notin \mathcal{N}_k. \end{cases} \quad (14)$$

Finally, dual updates are given by

$$y_{kj}^{t+1} = y_{kj}^t + \rho_k \{x_{kj}^{t+1} - z_j^{t+1}\} \quad j \in \mathcal{N}_k \quad (15)$$

Algorithm 1 summarizes the implementation of the distributed ADMM described here. The main feature of Algorithm 1 is that it does not require a master node, and all message passing is limited to the neighboring nodes only. The distributed implementation also requires that each node must be capable of carrying out the updates in (9) and (12). Algorithm 1 may be viewed as the proximal variant of the distributed ADMM algorithm [8], applied to the non-convex problem (2). The stopping criterion in Algorithm 1 is simply  $|z_k^{t+1} - z_k^t| \leq \delta$  for all  $k \in \mathcal{K}$  and a small  $\delta > 0$ .

Algorithm 1 is a synchronous protocol since all updates must necessarily be carried out at every iteration by every node. Its applicability to heterogeneous networks is therefore limited, since the progress of the algorithm is determined by the slowest node in the network. For instance, the following issues may arise when Algorithm 1 is implemented on a wireless network with energy-constrained, low-cost devices.

- (S1) For some nodes, calculating  $\nabla g_k(\cdot)$  [cf. step 8] or  $\text{prox}_k(\cdot)$  [cf. step 5] may be computationally demanding. In such cases, all nodes in the network must wait for the slowest node to carry out its update.
- (S2) Each node is required to transmit two messages to each of its neighbors per-iteration. This might be excessive for nodes operating on a power-budget.

---

**Algorithm 1 : Distributed Algorithm**


---

- 1: Initialize  $\{x_{kj}^1, y_{kj}^1\}$ ,  $\mathbf{z}_k$  for all  $j \in \mathcal{N}_k$ .
  - 2: **for**  $t = 1, 2, \dots$  **do**
  - 3:   Send  $\{\rho_k x_{kj}^t + y_{kj}^t\}$  to neighbors  $j \in \mathcal{N}_k$
  - 4:   Upon receiving  $\{\rho_j x_{jk}^t + y_{jk}^t\}$  for all  $j \in \mathcal{N}_k$ ,
  - 5:     Update  $z_k^{t+1}$  as in (9)
  - 6:   Transmit  $z_k^{t+1}$  to its neighbors  $j \in \mathcal{N}_k$
  - 7:   Upon receiving  $z_j^{t+1}$ , from all neighbors  $j \in \mathcal{N}_k$ ,
  - 8:     Update the primal variable  $\mathbf{x}_k^{t+1}$  as in (14)
  - 9:     Update the dual variable  $y_{kj}^{t+1}$  as in (15)
  - 10:   **if** stopping criteria is met **then**
  - 11:     terminate loop
  - 12:   **end if**
  - 13: **end for**
- 

Delays in function computation or communication may also arise because of the heterogeneity of nodes in a wireless network. For instance, mission critical nodes may attempt to extend their battery lives by operating under a power-saving mode, thus deliberately reducing their computational capabilities. On the other hand, the available power at some energy-harvesting nodes may vary throughout the day, depending on, for instance, the received solar energy.

Sec. III-B describes an asynchronous version of Algorithm 1 that overcomes (S1)-(S2) by allowing nodes to carry out the resource-intensive operations in Steps 5 and/or 7 intermittently. Each time an update is skipped, the node saves on both computational and communication costs. It is shown that the asynchronous algorithm converges, as long as the updates are performed “often enough,” a notion that is made precise through some constraints on the algorithm parameters.

### B. Distributed Asynchronous Algorithm with Optional and Delayed Updates

This section describes the asynchronous distributed ADMM algorithm that allows nodes to skip the calculation of  $\nabla g_k(\cdot)$  [cf. step 8] or  $\text{prox}_k(\cdot)$  [cf. step 5] for some iterations. For step 5, this is achieved by simply setting  $z_j^{t+1} = z_j^t$  for each non-updating node  $j \in \mathcal{K}$  at time  $t$ . Define  $\mathcal{S}^t$  as the set of nodes that carry out the update at time  $t$ , then Step 5 becomes

$$z_j^{t+1} = \begin{cases} \text{prox}_j \left( \frac{\sum_{k \in \mathcal{N}_j} (\rho_k x_{kj}^t + y_{kj}^t)}{\sum_{k \in \mathcal{N}_j} \rho_k} \right) & j \in \mathcal{S}^t \\ z_j^t & j \notin \mathcal{S}^t \end{cases} \quad (16)$$

Whenever  $j \notin \mathcal{S}^t$ , the subsequent transmission may not be carried out either, and the non-updating node may simply stay silent. The neighboring nodes will then wait for a fixed amount of time to receive an update, and assume that  $z_j^{t+1} = z_j^t$  holds for all nodes  $j \in \mathcal{N}_k$  that do not transmit anything.

For step 7, each node uses the latest available gradient  $\nabla g_k(\mathbf{z}_k^{[t+1]_k})$  for update in (14), where  $t+1 - T_k \leq [t+1]_k \leq t+1$  for some  $T_k < \infty$ . In other words, the gradient calculation may only be carried out intermittently, and the same gradient can be used for next several time slots. Alternatively, for computationally challenged nodes, the gradient calculation itself may take several time slots. Dropping the subscript  $k$



from  $[t+1]$  for notational brevity, the update in Step 7 becomes

$$x_{kj}^{t+1} = \begin{cases} z_j^{t+1} - \frac{1}{\rho_k} \left( [\nabla g_k(\mathbf{z}_k^{[t+1]})]_j + y_{kj}^t \right) & j \in \mathcal{N}_k \\ 0 & j \notin \mathcal{N}_k. \end{cases} \quad (17)$$

The transmission of updates in Step 3 is again optional and in the event that no update is received at a neighbor  $j \in \mathcal{N}_k$ , (16) is used instead.

---

**Algorithm 2** Distributed Asynchronous ADMM with Optional Updates

---

- 1: Set  $t = 1$ , initialize  $\{x_{kj}^1, y_{kj}^1, z_j^1\}$  for all  $j \in \mathcal{N}_k$ .
  - 2: **for**  $t = 1, 2, \dots$  **do**
  - 3:   (Optional) Send  $\{\rho_k x_{kj}^t + y_{kj}^t\}$  to neighbors  $j \in \mathcal{N}_k$
  - 4:   **if**  $\{\rho_j x_{jk}^t + y_{jk}^t\}$  received from all  $j \in \mathcal{N}_k$  **then**
  - 5:     (Optional) Update  $z_k^{t+1}$  as in (9) and transmit to each  $j \in \mathcal{N}_k$
  - 6:   **end if**
  - 7:   **if**  $z_j^{t+1}$  not received from some  $j \in \mathcal{N}_k$  **then**
  - 8:     set  $z_j^{t+1} = z_j^t$
  - 9:   **end if**
  - 10:   (Optional) Calculate gradient  $\nabla g_k(\mathbf{z}_k^{t+1})$
  - 11:   Update the primal variable  $\mathbf{x}_k^{t+1}$  as in (17)
  - 12:   Update the dual variable  $y_{kj}^{t+1}$  as in (15)
  - 13:   **if**  $\|\mathbf{x}_k^{t+1} - \mathbf{x}_k^t\| \leq \delta$  **then**
  - 14:     terminate loop
  - 15:   **end if**
  - 16: **end for**
- 

The proposed asynchronous algorithm for node  $k \in \mathcal{K}$  is summarized in Algorithm 2. The salient features of the proposed asynchronous algorithm are as follows.

- 1) All resource-intensive steps, such as gradient calculation, proximal function calculation, and transmission of updates are now optional.
- 2) Nodes operating on a power budget may only carry out the updates in Steps 11 and 12 at every iteration. When using the old gradient, these updates amount to simple addition/subtraction operations. It is also possible to defer Steps 11 and 12 to a later time slot when the transmission in Step 3 occurs.
- 3) The nodes may implement a timeout mechanism when listening for updates [cf. Steps 3 and 5]. The appropriate default actions must be triggered if nothing is heard from one or more neighbors.

Fig. 3 shows an illustration where nodes 2 and 3 are neighbors of a node 1, that go to sleep at iteration  $t = 2$ . During the sleep state, a node is only able to receive updates, but not transmit or carry out any computations. Observe that in the second sub-slot of the 2-nd iteration,  $z_2^3$  and  $z_3^3$  are also not updated at nodes 2 and 3 since no update for  $x_{12}^2$  and  $y_{12}^2$  was received. Consequently, it will hold that  $z_k^t = z_k^{t-1}$  for  $k = 1, 2$ , and 3. More generally, each sleeping node will force all its neighbors to use the update in Step 8 of Algorithm 2. Since the update frequency at each node cannot be too small, it is required that several nodes should be awake at any given time.

The next section provides the convergence analysis of

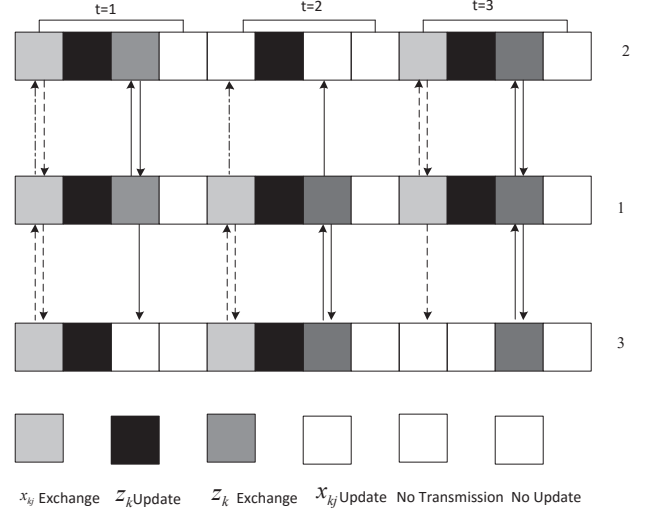


Fig. 3: Timing Diagram

Algorithm 2, ensuring that for any  $\delta > 0$ , the stopping criteria in Step 13 of Algorithm 2 is eventually met. Note that the convergence results also apply to Algorithm 1, which is simply a special case of Algorithm 2. Before concluding this subsection, the following remark about related work in the context of asynchronous algorithms is due.

**Remark 1.** The key feature of the proposed algorithms is that some of the updates and transmissions are entirely optional. From the perspective of algorithm design, this results in two sources of asynchrony, namely, delayed gradients, and skipped updates. The asynchronous algorithm in [16] can handle bounded delays in the gradient calculation, but does not consider skipping updates. This is because the DA-ADMM algorithm in [16] utilizes a fusion center that is not energy constrained. Note that the absence of such a fusion center significantly complicates the algorithm design, also restricting the application of Algorithms 2 to (2), which is a special case of (1).

An asynchronous ADMM algorithm was also proposed and applied to non-convex problem in [25]. However, no proof of convergence was provided, and the performance of the algorithm was only tested via simulations. A number of asynchronous variants exist for convex problems. The optional-update idea used in (16) is in fact inspired from the asynchronous ADMM proposed in [15], [28]. Different from Algorithm 2 however, the asynchronous algorithm in [28] also uses random network states, and is applicable only to convex problems. For partially separable problems such as (2), it may also be possible to develop an asynchronous variant of the coordinate descent algorithm. The stochastic asynchronous coordinate descent algorithm proposed in [22] also allows erroneous gradients but is applicable only to convex problems. Finally, asymptotic results on the effect of asynchrony on the stochastic gradient descent algorithm were provided in [35].

### C. Convergence Analysis for Algorithm 2

In order to establish the convergence of the asynchronous algorithm, some assumptions regarding the problem structure

and algorithm parameters must be made. In particular, it is shown that the algorithm converges as long as the optional updates happen “often enough.” Specifically, recall that  $t + 1 - \lfloor t + 1 \rfloor \leq T_k$ , which implies that the gradients may be calculated using updates that are at most  $T_k$ -old. Similarly, let the frequency of update in (9) being carried out at node  $k$  be denoted by  $0 < f_k \leq 1$ . For instance, if the update occurs once every  $K$  time slots,  $f_k = 1/K$ . Then the following assumptions are required.

**Assumption A1.** For each node  $k$ , the component function gradient  $\nabla g_k(\mathbf{x})$  is Lipschitz continuous, that is, there exists  $L_k > 0$ , for all  $\mathbf{x}, \mathbf{x}' \in \text{dom} g_k$  such that

$$\|\nabla g_k(\mathbf{x}) - \nabla g_k(\mathbf{x}')\| \leq L_k \|\mathbf{x} - \mathbf{x}'\|. \quad (18)$$

**Assumption A2.** The set  $\mathcal{X}$  is a closed, convex, and compact. The functions  $g_k(x)$  is bounded from below over  $\mathcal{X}$ .

**Assumption A3.** For node  $k$ , the step size  $\rho_k$  is chosen large enough such that, it holds that  $\alpha_k > 0$  and  $\beta_k > 0$ , where

$$\alpha_k := \frac{\rho_k f_k}{2} - \left( \frac{7L_k}{2\rho_k^2} + \frac{1}{\rho_k} \right) |\mathcal{N}_k| L_k^2 (T_k + 1)^2 - \frac{|\mathcal{N}_k| L_k T_k^2}{2} \quad (19)$$

$$\beta_k := \rho_k - 7L_k$$

Of these, Assumptions (A1) and (A2) are standard in the context of non-convex optimization [16], [17], [36] and are satisfied for most problems of interest. Intuitively, the use of old gradients is permitted only because they are Lipschitz continuous, and therefore change slowly over iterations. The boundedness assumption (A2) is required to ensure that the updates  $\{\mathbf{x}^t, \mathbf{z}^t, \mathbf{y}^t\}$  generated by Algorithm 2 stay bounded. This assumption may be dropped for certain problems where boundedness of these iterates may arise naturally. Finally, Assumption (A3) specifies the exact relationship that the algorithm and problem parameters  $\{L_k, \rho_k, f_k, T_k\}_{k=1}^K$  must satisfy in the worst case. Interestingly, the choice of  $\rho_k$  is node-specific since nodes may differ with respect to their computational capabilities resulting in different levels of asynchrony. While not stated explicitly, the convergence also requires that  $\rho_k < \infty$ . Thus, for (19) to hold, it is necessary that  $f_k > 0$  and  $T_k < \infty$ . In other words, the delay cannot be unbounded in the worst case.

The convergence of the asynchronous algorithm is established through the following intermediate lemma that holds under Assumptions (A1)-(A3).

**Lemma 1.** (a) Starting from any time  $t = t_0$ , there exists  $T < \infty$  such that

$$\begin{aligned} & L(\{\mathbf{x}_k^{T+t_0}\}; \mathbf{z}^{T+t_0}, \{\mathbf{y}_k^{T+t_0}\}) - L(\{\mathbf{x}_k^{t_0}\}; \mathbf{z}^{t_0}, \{\mathbf{y}_k^{t_0}\}) \\ & \leq - \sum_{i=t_0}^{T+t_0-1} \sum_{k=1}^K \frac{\beta_k}{2} \sum_{j \in \mathcal{N}_k} \|x_{kj}^{i+1} - x_{kj}^i\|^2 \\ & \quad - \sum_{i=t_0}^{T+t_0} \sum_{k=1}^K \alpha_k \sum_{j \in \mathcal{N}_k} \|z_j^{i+1} - z_j^i\|^2. \end{aligned} \quad (20)$$

(b) The augmented Lagrangian values in (8) are bounded from below, i.e., for any time  $t \geq 1$ , it holds that Lagrangian satisfies

$$L(\{\mathbf{x}_k^t\}; \mathbf{z}^t, \{\mathbf{y}_k^t\}) \geq P - \frac{L_k}{2} \sum_{j \in \mathcal{N}_k} \text{diam}^2(\mathcal{X}) > -\infty$$

The proof of Lemma 1 is provided in Appendix A. Lemma 1(a) establishes that there exists some finite  $T$  such that the augmented Lagrangian values are non-increasing after  $T$  iterations. In practice, the value of  $T$  depends on the update frequencies  $\{f_k\}_{k=1}^K$ . For instance,  $T$  could be the minimum number of iterations in which each node  $k$  updates  $z_k^t$  at least  $f_k T$  times. Lemma 1(b) establishes that the Lagrangian is bounded from below. One way to interpret Lemma 1 is to define the sequence  $L^n := L(\{\mathbf{x}_k^{nT+1}\}; \mathbf{z}^{nT+1}, \{\mathbf{y}_k^{nT+1}\})$  for all  $n \geq 0$ , and observe that  $\{L^n\}$  is non-increasing and bounded from below, and therefore convergent. The subsequent theorem establishes the final convergence result and related properties.

**Theorem 1.** (a) The iterates generated by Algorithm 2 converges in the following sense

$$\lim_{t \rightarrow \infty} \|z_k^{t+1} - z_k^t\| = 0, \quad \forall k \quad (21a)$$

$$\lim_{t \rightarrow \infty} \|x_{kj}^{t+1} - x_{kj}^t\| = 0, \quad j \in \mathcal{N}_k, \quad \forall k \quad (21b)$$

$$\lim_{t \rightarrow \infty} \|y_{kj}^{t+1} - y_{kj}^t\| = 0, \quad j \in \mathcal{N}_k, \quad \forall k \quad (21c)$$

(b) For each  $k \in \mathcal{K}$  and  $j \in \mathcal{N}_k$ , denote limit points of the sequences  $\{z_k^t\}$ ,  $\{x_{kj}^t\}$ , and  $\{y_{kj}^t\}$  by  $z_k^*$ ,  $x_{kj}^*$ , and  $y_{kj}^*$ , respectively. Then  $\{\{z_k^*\}, \{x_{kj}^*\}, \{y_{kj}^*\}\}$  is a stationary point of (5) and satisfies

$$\nabla g_k(\mathbf{x}_k^*) + \mathbf{y}_k^* = 0, \quad k = 1, \dots, K \quad (22a)$$

$$\sum_{j \in \mathcal{N}_k} y_{kj}^* \in \partial(h_k(z)) \mid_{z=z_k^*} \quad k = 1, \dots, K \quad (22b)$$

$$x_{kj}^* = z_j^* \in \mathcal{X}_j, \quad j \in \mathcal{N}_k, k = 1, \dots, K \quad (22c)$$

The proof of Theorem 1 is provided in Appendix B. Note that it suffices to show that Algorithm 2 converges to a stationary solution of (5) which is equivalent to (2). In other words,  $\{z_n^*\}_{n=1}^N$  can be used as a solution to (2). It is emphasized that Algorithm 2 may not necessarily converge to a globally optimum solution to (2).

Further, using assumptions (A1)-(A3) alone, it is also difficult to quantify the convergence rate of the ADMM algorithm for the non-convex case. As will be shown in Sec. V however, rate of convergence for the localization problem is low whenever  $\rho_k$  is large. Assuming that this observation applies to (2) generally, it makes sense to choose  $\rho_k$  as small as possible, while respecting (19). Interestingly, this result also matches with the intuition that the Algorithm 2 converges slowly if the asynchrony is high, i.e., when  $T_k \gg 1$  and  $f_k \ll 1$ , since it would require choosing a larger  $\rho_k$  for each  $k \in \mathcal{K}$ .

#### IV. MAJORIZED ASYNCHRONOUS ADMM

In many problems, it is possible to upper bound the non-convex component functions  $g_k(\mathbf{x}_k)$  with an appropriate convex surrogate function. Given  $\mathbf{z}_k \in \mathcal{X}$ , the surrogate  $f_k(\mathbf{x}_k, \mathbf{z}_k)$ , also referred to as the majorizing function, is such that for all  $\mathbf{x}_k$ ,  $g_k(\mathbf{x}_k) \leq f_k(\mathbf{x}_k, \mathbf{z}_k)$ , and satisfies

$$f_k(\mathbf{z}_k, \mathbf{z}_k) = g_k(\mathbf{z}_k) \quad (23)$$

$$\nabla f_k(\mathbf{z}_k, \mathbf{z}_k) := \nabla_{\mathbf{x}} f_k(\mathbf{x}, \mathbf{z}_k) \mid_{\mathbf{x}=\mathbf{z}_k} = \nabla_{\mathbf{x}} g_k(\mathbf{x}) \mid_{\mathbf{x}=\mathbf{z}_k}. \quad (24)$$

When such a majorizer exists and can be found easily, the update of  $\mathbf{x}_k^{t+1}$  in (14) can be carried out more accurately, without appreciable increase in the per-iteration complexity.

This section develops a provably convergent variant of Algorithm 2 that utilizes majorization while updating  $\mathbf{x}_k^{t+1}$ . Specifically, while (15) and (16) stay the same, the update for  $\mathbf{x}_k^{t+1}$  becomes

$$\mathbf{x}_k^{t+1} = \arg \min f_k(\mathbf{x}_k, \mathbf{z}_k^{[t+1]}) + \sum_{j \in \mathcal{N}_k} \langle y_{kj}^t, x_{kj} - z_j^{t+1} \rangle + \frac{\rho_k}{2} \sum_{j \in \mathcal{N}_k} \|x_{kj} - z_j^{t+1}\|^2 \quad (25)$$

where, as in (17),  $\mathbf{z}_k^{[t+1]}$  is used for calculating the surrogate function. Interestingly, the majorized ADMM proposed here enjoys the flexibility afforded by its proximal counterpart, and can be implemented as in Algorithms 2 or ??.

In order to show that the majorized ADMM converges, the following assumptions are required in addition to (A2).

**Assumption A4.** For each node  $k$ , there exists a constant  $L_k \geq 0$ , such that for all  $\mathbf{x}, \mathbf{x}', \mathbf{z}, \mathbf{z}'$ , the following inequalities are satisfied:

$$\|\nabla g_k(\mathbf{x}) - \nabla g_k(\mathbf{x}')\| \leq L_k \|\mathbf{x} - \mathbf{x}'\| \quad (26a)$$

$$\|\nabla f(\mathbf{x}, \mathbf{z}) - \nabla f(\mathbf{x}', \mathbf{z})\| \leq L_k \|\mathbf{x} - \mathbf{x}'\| \quad (26b)$$

$$\|\nabla f(\mathbf{x}, \mathbf{z}) - \nabla f(\mathbf{x}, \mathbf{z}')\| \leq L_k \|\mathbf{z} - \mathbf{z}'\|. \quad (26c)$$

**Assumption A5.** For node  $k$ , the step size  $\rho_k$  is chosen large enough such that  $\alpha_k > 0$  and  $\beta_k > 0$ , where

$$\alpha_k := |\mathcal{N}_k| \left[ \left( \frac{\rho_k f_k}{2} - \left( \frac{8L_k}{\rho_k^2} + \frac{1}{\rho_k} \right) L_k^2 (T_k + 1)^2 - \frac{L_k T_k^2}{2} \right) \right] \quad (27)$$

$$\beta_k := \frac{\rho_k - 9L_k}{2} - \frac{8L_k^3}{\rho_k^2}.$$

Note that Assumption (A4) utilizes the same Lipschitz constant in (26) for the sake of simplicity. In general, if there exist constants  $L_k^g$ ,  $L_k^x$ , and  $L_k^z$  in (26), it is always possible to define  $L_k := \max\{L_k^g, L_k^x, L_k^z\}$  for all  $k$ . Likewise, the convergence proof provided here can also be developed for the case when the three constants are different, resulting in slightly tighter bounds.

The conditions required in (A4) are not very restrictive. Consider for instance a component function  $g_k(\mathbf{x})$  that is expressible as a sum of a convex function  $g_k^1(\mathbf{x})$  and a concave function  $g_k^2(\mathbf{x})$ . The concavity of  $g_k^2(\mathbf{x})$  allows it to be majorized by its supporting hyperplane, i.e., given any  $\mathbf{z} \in \text{dom } g_k^2(\mathbf{x})$ ,

$$g_k(\mathbf{x}) = g_k^1(\mathbf{x}) + g_k^2(\mathbf{x}) \leq f_k(\mathbf{x}, \mathbf{z}) := g_k^1(\mathbf{x}) + g_k^2(\mathbf{z}) + \langle \nabla g_k^2(\mathbf{z}), \mathbf{x} - \mathbf{z} \rangle \quad (28)$$

where it can be verified that  $f_k(\mathbf{x}, \mathbf{z})$  satisfies (23) and (24). Then, observe that  $\nabla f_k(\mathbf{x}, \mathbf{z}) = \nabla g_k^1(\mathbf{x}) + \nabla g_k^2(\mathbf{x})$  for all  $\mathbf{x}$  and  $\mathbf{z}$ . If  $L_k^1$  and  $L_k^2$  are Lipschitz constants of  $\nabla g_k^1$  and  $\nabla g_k^2$  respectively, it may be seen that (26a) and (26b) hold with Lipschitz constant  $L_k^1 + L_k^2$ , while the left-hand side of (26c) is identically zero. The following Theorem, whose proof is provided in Appendix C, summarizes the main result of this section.

**Theorem 2.** The iterates generated by the majorized ADMM ((16), (25), and (15)) converge to a stationary point of (5).

## V. DISTRIBUTED LOCALIZATION IN NETWORKS

This section builds upon the localization examples introduced in Sec. II-C and provides simulation results for comparing two different related algorithms. Monte-Carlo simulations are performed over randomly generated networks with  $N = 25$  nodes. To this end, nodes are uniformly distributed over a unit two-dimensional area  $\mathcal{R} = [0, 1]^2$ . Of these,  $m = 5$  nodes are located at  $(0.25, 0.25)$ ,  $(0.75, 0.25)$ ,  $(0.25, 0.75)$ ,  $(0.5, 0.5)$ , at  $(0.75, 0.75)$ , and serve as anchor nodes for others. Distance measurement and communication between nodes  $k$  and  $j$  is possible only if they are within a distance of  $R = 0.5$  of each other. For neighboring nodes, the weights  $w_{kj}$  are all set to unity.

It is remarked that the classical SMACOF algorithm used for solving in [37] is not distributed, and is therefore not applicable in the present context. The distributed weighted MDS (DwMDS) framework introduced in [4] is an incremental algorithm that utilizes component wise majorization in order to circumvent the non-convexity of the objective function. Different from the proposed algorithms, incremental algorithms utilize a cyclic message passing routine and are synchronous in nature.

A distributed and asynchronous algorithm utilizing semidefinite relaxation (SDR) was proposed in [3]. The so-called E-ML (edge relaxed maximum likelihood) algorithm utilizes relaxation to first convert the objective into a convex one, and then applies distributed ADMM. Since ADMM is applied on a convex and relaxed version of (4), it can be made distributed and asynchronous using known techniques. Specifically, E-ML activates only a subset of edges  $\mathcal{E}^{(t)} \subset \mathcal{E}$  at each iteration, allowing nodes to save energy by not transmitting on some links [38]. Different from the present Algorithms 2, which converges to a stationary point of (4), the E-ML algorithm converges to the global optimum of the relaxed version of (4). Since both limit points may be different from the global optimum of (4) itself, the subsequent comparisons utilize the normalized root mean square error (NRMSE) of the location estimates, given

by  $NRMSE := \sqrt{\mathbb{E} [\|\hat{\mathbf{X}} - \mathbf{X}^*\|_F^2] / \mathbb{E} [\|\mathbf{X}^*\|_F^2]}$  where the expectations are performed via Monte-Carlo simulations using 100 independent realizations of the random network.

In the present case, the following modified version of the objective function in (4) is considered  $d_{kj}(\mathbf{x}_k, \mathbf{x}_j) = \sqrt{\|\mathbf{x}_k - \mathbf{x}_j\|} + \epsilon$ , where  $\epsilon > 0$  is a small number introduced to make the objective function differentiable everywhere. This modification also makes  $\nabla g_k$  Lipschitz continuous, as required in (A1), and verified in Appendix D. Recall that Algorithm 2 allows two modes of asynchrony, namely, old gradient and skipped updates. In the implementation described here, it is assumed that at each time instant, some nodes go to sleep and do not carry out any update. Further, at each iteration, each node randomly chooses an older available gradient or calculates a more recent one, while ensuring that the gradient used is at most  $T_k$ -old.

### A. NRMSE Performance

Fig. 4 compares the NRMSE performance of DwMDS algorithm [4], the synchronous (SyncE-ML) and asynchronous

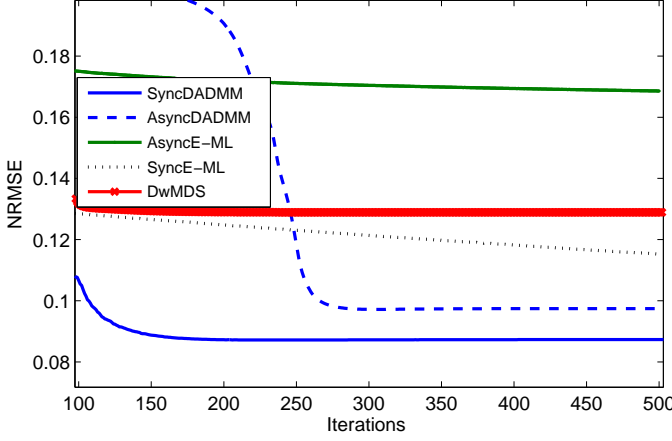


Fig. 4: NRMSE performance

(AsyncE-ML) versions of the E-ML algorithm [3], and the Algorithms 1 (SyncDADMM), 2 (AsyncDADMM). For the asynchronous E-ML algorithm, 125 of the 130 edges are activated per iteration. For Algorithm 2, we set  $T_k = 8$  and  $f_k = 0.75$ , while  $\rho_k$  is chosen so as to satisfy (A3).

From Fig. 4, it is clear that both, Algorithms 1, 2 outperform all the other state-of-the-art algorithms. As expected, the asynchronous versions of the E-ML and ADMM algorithms have poorer performance than their synchronous counterparts. Interestingly, asynchrony has a greater effect on the asymptotic performance of the E-ML algorithm than the convergence rate of the ADMM algorithm. It is also worthwhile to compare the implementation complexity of the two asynchronous algorithms. Since the E-ML algorithm uses SDR, its computational complexity is approximately  $\mathcal{O}(|\mathcal{N}_k|^3)$  for node  $k$  per iteration. This is because each iteration requires solving an  $|\mathcal{N}_k|$ -sized convex semidefinite program. On the other hand, gradient calculation is the most computationally demanding step in Algorithms 1 and 2, which translates to a total computational complexity of  $\mathcal{O}(|\mathcal{N}_k|)$  for node  $k$  per iteration. Note further that unlike Algorithm 2, the updates in AsyncE-ML algorithm are also not optional, irrespective of the number of edges activated per iteration. The communication complexity for the two algorithms is of the same order, i.e.,  $\mathcal{O}(|\mathcal{N}_k|)$  for node  $k$  per iteration. However, the message transmitted to each neighbor in AsyncE-ML consists of nine real numbers, as opposed to four real numbers in Algorithm 2.

### B. Convergence rates

As stated earlier, while Theorem 1 establishes convergence, it does not provide any indication regarding the convergence rate exhibited by Algorithm 2. This subsection compares the convergence rates of the synchronous proximal ADMM (SyncADMM), synchronous majorized ADMM (SyncMDADMM), and the consensus-based distributed gradient descent (C-DGD) method proposed in [13, Chap. 10]. While C-DGD has only been proposed for convex problems, it has been implemented in a distributed and asynchronous manner [39], [40], and is therefore an interesting candidate for the non-convex localization problem. The goal is to compare the three algorithms

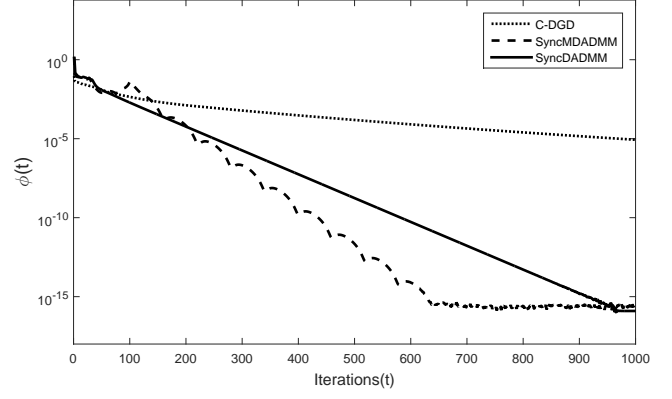


Fig. 5: Convergence of Majorized and Proximal Variant and C-DGD

in terms of how fast they approach a local minimum, while ignoring their distances to the global minimum. To this end, Fig. 5 shows the evolution of the convergence criterion  $\phi(t) := \left\| \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i^{t+1} - \mathbf{x}_i^t) \right\|_F$  with the iteration index  $t$ . It can be observed that the majorized ADMM performs slightly better than the proximal ADMM. On the other hand, the C-DGD algorithm convergence relatively slowly. In particular, unlike the other simulations, the plots shown in Fig. 5 are generated for a network with communication range  $R = 0.8$ , since the C-DGD algorithm does not converge for  $R = 0.5$ . It is remarked that as the range decreases, fewer measurements are available, making network localization more difficult.

At this point, it may be useful to also compare the complexity and run-times of the three algorithms. In general, the majorized ADMM is the most complex, since it involves solving a convex optimization problem at every iteration. On the other hand, both the C-DGD and the proximal ADMM algorithms require simple calculations with the gradients of  $g_k$ . For the localization problem considered here, the per-iteration majorization update at node  $k$  requires solving a system of  $|\mathcal{N}_k| \times |\mathcal{N}_k|$  linear equations, incurring almost four times as much CPU-time as the proximal ADMM and the C-DGD algorithms. This extra complexity more than compensates for the reduced number of iterations afforded by the majorized ADMM. In summary, the proximal ADMM algorithm takes the least wall-time to reach a certain accuracy, say  $\phi(t) = 10^{-15}$ .

### C. Choice of parameters

In order to further study the convergence rates, performance of Algorithm 2 is studied for different parameter values. The convergence rate is analyzed by plotting the stopping criterion given by  $\psi(t) := \|\mathbf{z}^{t+1} - \mathbf{z}^t\|$  against the iteration index  $t$ .

1) *Choice of  $\rho_k$* : Theorem 1 guarantees the convergence of Algorithm 2 whenever  $\rho_k$  is chosen in accordance with (A3). In practice however, it may be possible to improve the convergence rate of Algorithm 2 by choosing smaller values of  $\rho_k$ . Fig. 6 shows the evolution of  $\psi(t)$  and NRMSE with iterations, for different values of  $\rho_k$ , while  $T_k = 4$  and  $f_k = 0.75$  are kept fixed. As expected, the algorithm takes longer



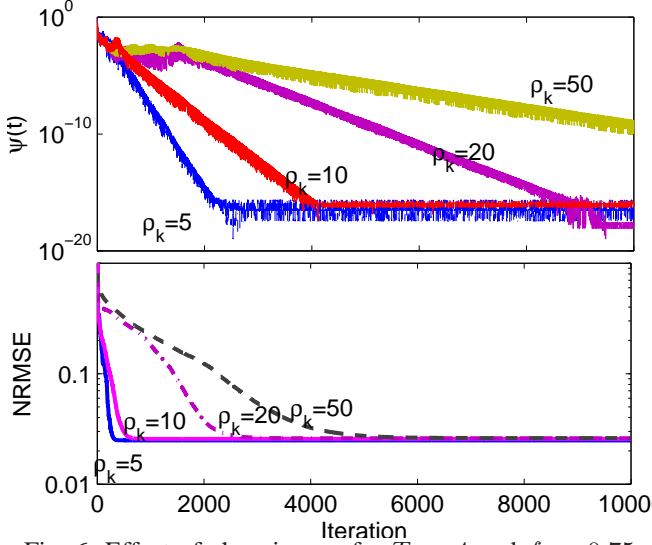


Fig. 6: Effect of choosing  $\rho_k$  for  $T_k = 4$  and  $f_k = 0.75$

to converge for larger values of  $\rho_k$ , although the asymptotic NRMSE performance for all cases remains the same.

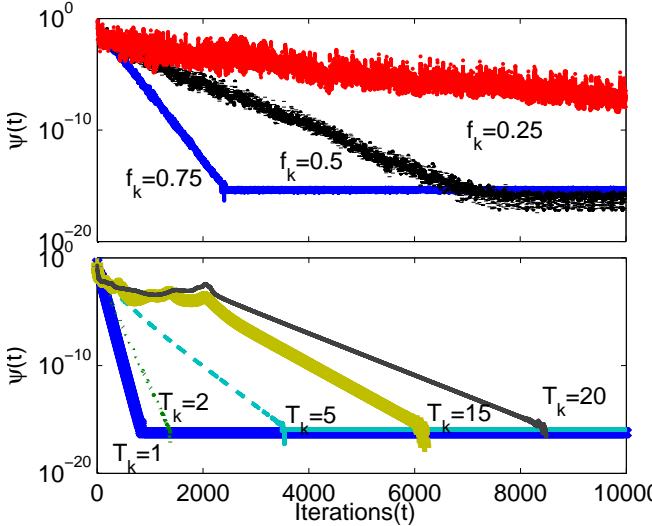


Fig. 7: Effect of asynchrony for  $\rho_k = 10$ . The top figure uses  $T_k = 4$ , while the bottom figure uses  $f_k = 0.75$ .

2) *Effect of Asynchrony*: Fig. 7 shows the effect of choosing the parameters  $T_k$  and  $f_k$  on the rate of convergence. To this end, we set  $\rho_k = 10$  for all  $k \in \mathcal{N}$ , and vary  $T_k$  and  $f_k$  separately. Both figures confirm the intuition that introduction of asynchrony results in slower convergence, even when  $\rho_k$  stays the same. Note however that in order to guarantee convergence, it is necessary to choose increasingly larger values of  $\rho_k$  for larger values of  $T_k$  or smaller values of  $f_k$ . Such a choice may therefore result in even slower convergence but allow higher asynchrony.

## VI. CONCLUSION

This paper develops an asynchronous distributed ADMM algorithm that is applicable to a class of non-convex optimization problems. The non-convexity of the cost functions

is handled either by making a first order approximation or via majorization, resulting in two variants of the proposed algorithm. Both variants converge to a stationary point of the optimization problem as long as the ADMM updates are applied “often enough.” The proposed algorithms find applications in distributed in-network estimation and localization. Comparisons with state-of-the-art distributed algorithms for the problem of cooperative localization in ad hoc networks demonstrates the superior performance of the proposed algorithm.

## APPENDIX A PROOF OF LEMMA 1

This appendix provides the convergence analysis for the Lagrangian in Algorithm 2. Before proceeding with the proof, some notation is introduced. Recall the definitions of  $\mathbf{x}_k$ ,  $\mathbf{y}_k$ , and  $\mathbf{z}_k$ , and similarly define  $[\tilde{\mathbf{x}}_j]_k := x_{kj}$  and  $[\tilde{\mathbf{y}}_j]_k := y_{kj}$  for all  $k \in \mathcal{N}_j$ . Since the Lagrangian is separable in both  $\{\mathbf{x}_k\}$  and  $\{\mathbf{z}_j\}$ , the following notation is introduced for the summands in (7) and (8):

$$\begin{aligned} \ell_k(\mathbf{x}_k, \mathbf{z}_k, \mathbf{y}_k) &:= g_k(\mathbf{x}_k) + \sum_{j \in \mathcal{N}_k} \langle y_{kj}, x_{kj} - z_j \rangle \\ &\quad + \sum_{j \in \mathcal{N}_k} \frac{\rho_k}{2} \|x_{kj} - z_j\|^2 \end{aligned} \quad (29a)$$

$$\begin{aligned} \tilde{\ell}_j(\tilde{\mathbf{x}}_j, z_j, \tilde{\mathbf{y}}_j) &:= \sum_{k \in \mathcal{N}_j} \langle y_{kj}, x_{kj} - z_j \rangle \\ &\quad + \sum_{k \in \mathcal{N}_j} \frac{\rho_k}{2} \|x_{kj} - z_j\|^2 + h_j(z_j) \end{aligned} \quad (29b)$$

so that  $L(\{\mathbf{x}_k\}, \mathbf{z}, \{\mathbf{y}_k\}) = \sum_k \ell_k(\mathbf{x}_k, \mathbf{z}_k, \mathbf{y}_k) + \sum_k h_k(z_k) = \sum_j g_j(\mathbf{x}_j) + \sum_j \tilde{\ell}_j(\tilde{\mathbf{x}}_j, z_j, \tilde{\mathbf{y}}_j)$ . Further, the approximation of  $\ell_k(\mathbf{x}_k, \mathbf{z}_k, \mathbf{y}_k)$  at  $\mathbf{x}_k = \tilde{\mathbf{z}}$  is denoted by [cf. (12)]

$$\begin{aligned} u_k(\mathbf{x}_k, \tilde{\mathbf{z}}, \mathbf{z}_k, \mathbf{y}_k) &:= g_k(\tilde{\mathbf{z}}) + \langle \nabla g_k(\tilde{\mathbf{z}}), \mathbf{x}_k - \tilde{\mathbf{z}} \rangle \\ &\quad + \sum_{j \in \mathcal{N}_k} \langle y_{kj}, x_{kj} - z_j \rangle + \sum_{j \in \mathcal{N}_k} \frac{\rho_k}{2} \|x_{kj} - z_j\|^2. \end{aligned} \quad (30)$$

With this definition, observe that the update for  $\mathbf{x}_k^{t+1}$  in (12) is given by

$$\mathbf{x}_k^{t+1} = \arg \min_{\mathbf{x}_k} u_k(\mathbf{x}_k, \mathbf{z}_k^{[t+1]}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t) \quad (31)$$

$$\Rightarrow \nabla_{\mathbf{x}_k} u_k(\mathbf{x}_k^{t+1}, \mathbf{z}_k^{[t+1]}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t) = 0 \quad (32)$$

where the gradient with respect to  $\mathbf{x}_k$  is defined similar to that in (13). In order to show that the Lagrangian decreases over several iterations, express the difference between consecutive Lagrangian values as

$$\begin{aligned} &L(\{\mathbf{x}_k^{t+1}\}, \mathbf{z}^{t+1}, \{\mathbf{y}_k^{t+1}\}) - L(\{\mathbf{x}_k^t\}, \mathbf{z}^t, \{\mathbf{y}_k^t\}) \\ &= L(\{\mathbf{x}_k^{t+1}\}, \mathbf{z}^{t+1}, \{\mathbf{y}_k^{t+1}\}) - L(\{\mathbf{x}_k^{t+1}\}, \mathbf{z}^{t+1}, \{\mathbf{y}_k^t\}) \\ &\quad + L(\{\mathbf{x}_k^{t+1}\}, \mathbf{z}^{t+1}, \{\mathbf{y}_k^t\}) - L(\{\mathbf{x}_k^t\}, \mathbf{z}^{t+1}, \{\mathbf{y}_k^t\}) \\ &\quad + L(\{\mathbf{x}_k^t\}, \mathbf{z}^{t+1}, \{\mathbf{y}_k^t\}) - L(\{\mathbf{x}_k^t\}, \mathbf{z}^t, \{\mathbf{y}_k^t\}). \end{aligned} \quad (33)$$

The subsequent lemma establishes bounds on the different terms in (33), and will be utilized to prove Lemma 1(a).

**Lemma 2.** Define  $S^t$  as the set of nodes for which the update in (16) is carried out at time  $t$ . Then it holds that

$$\begin{aligned} &L(\{\mathbf{x}_k^{t+1}\}, \mathbf{z}^{t+1}, \{\mathbf{y}_k^{t+1}\}) - L(\{\mathbf{x}_k^{t+1}\}, \mathbf{z}^{t+1}, \{\mathbf{y}_k^t\}) \\ &\leq \sum_{k=1}^K \frac{L_k^2(T_k + 1)}{\rho_k} \sum_{j \in \mathcal{N}_k} \sum_{d=0}^{T_k} \|z_j^{t+1-d} - z_j^{t-d}\|^2 \end{aligned} \quad (34)$$

$$\begin{aligned}
& L(\{\mathbf{x}_k^{t+1}\}, \mathbf{z}^{t+1}, \{\mathbf{y}_k^t\}) - L(\{\mathbf{x}_k^t\}, \mathbf{z}^{t+1}, \{\mathbf{y}_k^t\}) \\
& \leq -\sum_{k=1}^K \frac{\rho_k - 7L_k}{2} \sum_{j \in \mathcal{N}_k} \|x_{kj}^{t+1} - x_{kj}^t\|^2 \\
& + \sum_{k=1}^K \left( \frac{7L_k^3(T_k + 1)}{2\rho_k^2} + \frac{L_k T_k}{2} \right) \sum_{d=0}^{T_k-1} \sum_{j \in \mathcal{N}_k} \|z_j^{t+1-d} - z_j^{t-d}\|^2
\end{aligned} \tag{35}$$

$$\begin{aligned}
& L(\{\mathbf{x}_k^t\}, \mathbf{z}^{t+1}, \{\mathbf{y}_k^t\}) - L(\{\mathbf{x}_k^t\}, \mathbf{z}^t, \{\mathbf{y}_k^t\}) \\
& \leq -\sum_{k \in \mathcal{S}^t} \frac{\rho_k}{2} \|z_k^{t+1} - z_k^t\|^2
\end{aligned} \tag{36}$$

*Proof.* The proof begins by establishing a bound on difference between successive dual values. Rearranging (17),  $[\nabla g_k(\mathbf{z}_k^{[t+1]})]_j + y_{kj}^t + \rho_k(x_{kj}^{t+1} - z_j^{t+1}) = 0$ ,  $j \in \mathcal{N}_k$  which, together with the update for  $y_{kj}^{t+1}$  [cf. (15)] yields

$$[\nabla g_k(\mathbf{z}_k^{[t+1]})]_j = -y_{kj}^{t+1}, \quad j \in \mathcal{N}_k. \tag{37}$$

Similarly, it holds that  $y_{kj}^t = -[\nabla g_k(\mathbf{z}_k^{[t]})]_j$  for all  $j \in \mathcal{N}_k$ . Therefore, for each  $k \in \mathcal{K}$ , the following bound applies:

$$\begin{aligned}
\sum_{j \in \mathcal{N}_k} \|y_{kj}^{t+1} - y_{kj}^t\|^2 &= \|\nabla g_k(\mathbf{z}_k^{[t+1]}) - \nabla g_k(\mathbf{z}_k^{[t]})\|^2 \\
&\leq L_k^2 \|\mathbf{z}_k^{[t+1]} - \mathbf{z}_k^{[t]}\|^2
\end{aligned} \tag{38a}$$

$$\leq L_k^2 \left( \sum_{d=0}^{T_k} \|\mathbf{z}_k^{t+1-d} - \mathbf{z}_k^{t-d}\| \right)^2 \tag{38b}$$

$$\leq L_k^2 (T_k + 1) \sum_{d=0}^{T_k} \sum_{j \in \mathcal{N}_k} \|z_j^{t+1-d} - z_j^{t-d}\|^2 \tag{38c}$$

where (38a) follows from Assumption (A1) while (38b) and (38c) follow from the use of triangle inequality. Next, the different bounds in (34), (35), and (36) are proved.

(a) The bound in (34) follows from the following equalities utilizing (15).

$$\begin{aligned}
& L(\{\mathbf{x}_k^{t+1}\}, \mathbf{z}^{t+1}, \{\mathbf{y}_k^{t+1}\}) - L(\{\mathbf{x}_k^{t+1}\}, \mathbf{z}^{t+1}, \{\mathbf{y}_k^t\}) \\
&= \sum_{k=1}^K \sum_{j \in \mathcal{N}_k} \langle y_{kj}^{t+1}, x_{kj}^{t+1} - z_j^{t+1} \rangle - \langle y_{kj}^t, x_{kj}^{t+1} - z_j^{t+1} \rangle \\
&= \sum_{k=1}^K \sum_{j \in \mathcal{N}_k} \langle y_{kj}^{t+1} - y_{kj}^t, \frac{1}{\rho_k} \{y_{kj}^{t+1} - y_{kj}^t\} \rangle \\
&= \sum_{k=1}^K \frac{1}{\rho_k} \sum_{j \in \mathcal{N}_k} \|y_{kj}^{t+1} - y_{kj}^t\|^2
\end{aligned} \tag{39}$$

Finally, the bound in (34) follows by substituting (38c) into the right-hand side of (39).

(b) First some properties of the individual summands  $\ell_k(\mathbf{x}_k, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t)$  and their approximated versions  $u_k(\mathbf{x}_k, \tilde{\mathbf{z}}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t)$  are established. Since  $\nabla g_k(\mathbf{x}_k)$  is Lipschitz continuous in  $\mathbf{x}_k$ , the following quadratic upper bound holds for all  $\mathbf{x}_k$

$$\begin{aligned}
g_k(\mathbf{x}_k) &\leq g_k(\mathbf{z}_k^{t+1}) + \langle \nabla g_k(\mathbf{z}_k^{t+1}), \mathbf{x}_k - \mathbf{z}_k^{t+1} \rangle \\
&\quad + \frac{L_k}{2} \sum_{j \in \mathcal{N}_k} \|x_{kj} - z_j^{t+1}\|^2
\end{aligned} \tag{40}$$

which implies that

$$\ell_k(\mathbf{x}_k, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t) \leq u_k(\mathbf{x}_k, \mathbf{z}_k^{t+1}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t)$$

$$+ \frac{L_k}{2} \sum_{j \in \mathcal{N}_k} \|x_{kj} - z_j^{t+1}\|^2. \tag{41}$$

Observing that  $u_k(\mathbf{x}_k, \tilde{\mathbf{z}}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t)$  is differentiable and strongly convex with respect to  $\mathbf{x}_k$ , it holds that

$$\begin{aligned}
& u_k(\mathbf{x}_k^{t+1}, \tilde{\mathbf{z}}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t) - u_k(\mathbf{x}_k^t, \tilde{\mathbf{z}}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t) \\
& \leq \langle \nabla_{\mathbf{x}_k} u_k(\mathbf{x}_k^{t+1}, \tilde{\mathbf{z}}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t), \mathbf{x}_k^{t+1} - \mathbf{x}_k^t \rangle \\
& \quad - \frac{\rho_k}{2} \|\mathbf{x}_k^{t+1} - \mathbf{x}_k^t\|^2
\end{aligned} \tag{42}$$

for all  $\tilde{\mathbf{z}}$ . In particular for  $\tilde{\mathbf{z}} = \mathbf{z}_k^{[t+1]}$ , the gradient on the right is zero from (32), so that  $u_k(\mathbf{x}_k^{t+1}, \mathbf{z}_k^{[t+1]}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t) - u_k(\mathbf{x}_k^t, \mathbf{z}_k^{[t+1]}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t) \leq -\frac{\rho_k}{2} \sum_{j \in \mathcal{N}_k} \|x_{kj}^{t+1} - x_{kj}^t\|^2$ . Note also that  $\nabla_{\mathbf{x}_k} u_k(\mathbf{x}_k^{t+1}, \mathbf{z}_k^{[t+1]}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t) - \nabla_{\mathbf{x}_k} u_k(\mathbf{x}_k^{t+1}, \mathbf{z}_k^{t+1}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t) = \nabla g_k(\mathbf{z}_k^{[t+1]}) - \nabla g_k(\mathbf{z}_k^{t+1})$

Next, specializing (42) for  $\tilde{\mathbf{z}} = \mathbf{z}_k^{t+1}$ , the following series of inequalities are obtained

$$\begin{aligned}
& u_k(\mathbf{x}_k^{t+1}, \mathbf{z}_k^{t+1}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t) - u_k(\mathbf{x}_k^t, \mathbf{z}_k^{t+1}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t) \\
& \leq \langle \nabla_{\mathbf{x}_k} u_k(\mathbf{x}_k^{t+1}, \mathbf{z}_k^{t+1}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t), \mathbf{x}_k^{t+1} - \mathbf{x}_k^t \rangle \\
& \quad - \frac{\rho_k}{2} \|\mathbf{x}_k^{t+1} - \mathbf{x}_k^t\|^2
\end{aligned} \tag{43a}$$

$$\begin{aligned}
&= \langle \nabla_{\mathbf{x}_k} u_k(\mathbf{x}_k^{t+1}, \mathbf{z}_k^{t+1}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t), \mathbf{x}_k^{t+1} - \mathbf{x}_k^t \rangle \\
&\quad - \langle \nabla_{\mathbf{x}_k} u_k(\mathbf{x}_k^{[t+1]}, \mathbf{z}_k^{t+1}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t), \mathbf{x}_k^{t+1} - \mathbf{x}_k^t \rangle \\
&\quad - \frac{\rho_k}{2} \|\mathbf{x}_k^{t+1} - \mathbf{x}_k^t\|^2
\end{aligned} \tag{43b}$$

$$\begin{aligned}
&= \langle \nabla g_k(\mathbf{z}_k^{t+1}) - \nabla g_k(\mathbf{z}_k^{[t+1]}), \mathbf{x}_k^{t+1} - \mathbf{x}_k^t \rangle \\
&\quad - \frac{\rho_k}{2} \|\mathbf{x}_k^{t+1} - \mathbf{x}_k^t\|^2 \\
&\leq L_k \|\mathbf{z}_k^{[t+1]} - \mathbf{z}_k^{t+1}\| \|\mathbf{x}_k^{t+1} - \mathbf{x}_k^t\| - \frac{\rho_k}{2} \|\mathbf{x}_k^{t+1} - \mathbf{x}_k^t\|^2
\end{aligned} \tag{43c}$$

$$\begin{aligned}
&\leq \frac{L_k T_k}{2} \sum_{d=0}^{T_k-1} \|\mathbf{z}_k^{t+1-d} - \mathbf{z}_k^{t-d}\|^2 - \frac{\rho_k - L_k}{2} \|\mathbf{x}_k^{t+1} - \mathbf{x}_k^t\|^2 \\
&= \frac{L_k T_k}{2} \sum_{d=0}^{T_k-1} \sum_{j \in \mathcal{N}_k} \|z_j^{t+1-d} - z_j^{t-d}\|^2 \\
&\quad - \frac{\rho_k - L_k}{2} \sum_{j \in \mathcal{N}_k} \|x_{kj}^{t+1} - x_{kj}^t\|^2
\end{aligned} \tag{43d}$$

where (43b) follows from (32), (43c) follows from (A1), and the rest follow from the use of triangle inequality similar to that in (38).

Similarly, observe that

$$\begin{aligned}
& u_k(\mathbf{x}_k^{t+1}, \mathbf{z}_k^{t+1}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t) - \ell_k(\mathbf{x}_k^{t+1}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t) \\
&= g_k(\mathbf{z}_k^{t+1}) - g_k(\mathbf{x}_k^t) + \langle \nabla g_k(\mathbf{z}_k^{t+1}), \mathbf{x}_k^t - \mathbf{z}_k^{t+1} \rangle.
\end{aligned} \tag{44}$$

Similar to (40), the following quadratic upper bound is also implied by (A1),

$$\begin{aligned}
& g_k(\mathbf{z}_k^{t+1}) \leq g_k(\mathbf{x}_k^t) + \langle \nabla g_k(\mathbf{x}_k^t), \mathbf{z}_k^{t+1} - \mathbf{x}_k^t \rangle \\
&\quad + \frac{L_k}{2} \sum_{j \in \mathcal{N}_k} \|x_{kj}^t - z_j^{t+1}\|^2.
\end{aligned} \tag{45}$$

Substituting (45) into (44), we obtain

$$\begin{aligned}
& u_k(\mathbf{x}_k^{t+1}, \mathbf{z}_k^{t+1}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t) - \ell_k(\mathbf{x}_k^{t+1}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t) \\
& \leq \langle \nabla g_k(\mathbf{z}_k^{t+1}) - \nabla g_k(\mathbf{x}_k^t), \mathbf{x}_k^t - \mathbf{z}_k^{t+1} \rangle \\
& \quad + \frac{L_k}{2} \sum_{j \in \mathcal{N}_k} \|x_{kj}^t - z_j^{t+1}\|^2
\end{aligned}$$

$$\leq \frac{3L_k}{2} \sum_{j \in \mathcal{N}_k} \|x_{kj}^t - z_j^{t+1}\|^2 \quad (46a)$$

$$\leq 3L_k \sum_{j \in \mathcal{N}_k} \left( \|x_{kj}^t - x_{kj}^{t+1}\|^2 + \|x_{kj}^{t+1} - z_j^{t+1}\|^2 \right) \quad (46b)$$

where (46a) follows from (A1) and (46b) from the use of triangle inequality.

Having derived inequalities (41)-(46), the difference between the summands of (35) can finally be bounded. Towards this end, it holds that

$$\begin{aligned} & \ell_k(\mathbf{x}_k^{t+1}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t) - \ell_k(\mathbf{x}_k^t, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t) \\ & \leq u_k(\mathbf{x}_k^{t+1}, \mathbf{z}_k^{t+1}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t) - u_k(\mathbf{x}_k^t, \mathbf{z}_k^{t+1}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t) \\ & + \frac{L_k}{2} \sum_{j \in \mathcal{N}_k} \|z_j^{t+1} - x_{kj}^{t+1}\|^2 \\ & + u_k(\mathbf{x}_k^t, \mathbf{z}_k^{t+1}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t) - \ell_k(\mathbf{x}_k^t, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t) \\ & \leq \frac{L_k T_k}{2} \sum_{d=0}^{T_k-1} \sum_{j \in \mathcal{N}_k} \|z_j^{t+1-d} - z_j^{t-d}\|^2 \\ & - \frac{\rho_k - L_k}{2} \sum_{j \in \mathcal{N}_k} \|x_{kj}^{t+1} - x_{kj}^t\|^2 + \frac{L_k}{2} \sum_{j \in \mathcal{N}_k} \|z_j^{t+1} - x_{kj}^{t+1}\|^2 \\ & + 3L_k \sum_{j \in \mathcal{N}_k} \left( \|x_{kj}^t - x_{kj}^{t+1}\|^2 + \|x_{kj}^{t+1} - z_j^{t+1}\|^2 \right) \quad (47) \end{aligned}$$

$$\begin{aligned} & = -\frac{\rho_k - 7L_k}{2} \sum_{j \in \mathcal{N}_k} \|x_{kj}^{t+1} - x_{kj}^t\|^2 + \frac{7L_k}{2\rho_k^2} \sum_{j \in \mathcal{N}_k} \|y_{kj}^{t+1} - y_{kj}^t\|^2 \\ & + \frac{L_k T_k}{2} \sum_{d=0}^{T_k-1} \sum_{j \in \mathcal{N}_k} \|z_j^{t+1-d} - z_j^{t-d}\|^2 \quad (48) \\ & = -\frac{\rho_k - 7L_k}{2} \sum_{j \in \mathcal{N}_k} \|x_{kj}^{t+1} - x_{kj}^t\|^2 \\ & + \left( \frac{7L_k^3(T_k+1)}{2\rho_k^2} + \frac{L_k T_k}{2} \right) \sum_{d=0}^{T_k-1} \sum_{j \in \mathcal{N}_k} \|z_j^{t+1-d} - z_j^{t-d}\|^2 \quad (49) \end{aligned}$$

where (48) follows from using the update for  $y_{kj}^{t+1}$  and (49) from (38). Finally, summing (49) over  $k = 1, 2, \dots, K$ , the result in (35) follows.

(c) Define the indicator function  $\mathbf{1}_{\mathcal{X}}(x) = 1$  if  $x \in \mathcal{X}$  and zero otherwise, the observe that update for  $z_j$  can be written as [cf. (9)],

$$z_j^{t+1} = \arg \min_{z_j \in \mathcal{X}_j} \tilde{\ell}(\mathbf{x}_j^t, z_j, \mathbf{y}_j^t) \quad (50)$$

$$= \arg \min_{z_j} \tilde{\ell}(\mathbf{x}_j^t, z_j, \mathbf{y}_j^t) + \mathbf{1}_{\mathcal{X}_j}(z_j) \quad (51)$$

for all  $j \in \mathcal{S}^t$ . The first order optimality condition for (51) is that

$$0 \in \partial(\tilde{\ell}(\mathbf{x}_j^t, z_j^{t+1}, \mathbf{y}_j^t) + \mathbf{1}_{\mathcal{X}_j}(z_j^{t+1})) \quad (52)$$

where  $\partial(\cdot)$  denotes the subgradient operator. Observe further that since the function  $\tilde{\ell}(\mathbf{x}_j^t, z_j, \mathbf{y}_j^t) + \mathbf{1}_{\mathcal{X}_j}(z_j)$  is strongly convex in  $z_j$ , it holds that

$$\begin{aligned} & \tilde{\ell}(\mathbf{x}_j^t, z_j^{t+1}, \mathbf{y}_j^t) + \mathbf{1}_{\mathcal{X}_j}(z_j^{t+1}) \\ & \leq \tilde{\ell}(\mathbf{x}_j^t, z_j^t, \mathbf{y}_j^t) + \mathbf{1}_{\mathcal{X}_j}(z_j^t) - \frac{\rho_j}{2} \|z_j^{t+1} - z_j^t\|^2 \\ & - \langle \partial(\tilde{\ell}(\mathbf{x}_j^t, z_j^{t+1}, \mathbf{y}_j^t) + \mathbf{1}_{\mathcal{X}_j}(z_j^{t+1})), z_j^t - z_j^{t+1} \rangle. \quad (53) \end{aligned}$$

Since  $\mathbf{1}_{\mathcal{X}_j}(z_j^{t+1}) = \mathbf{1}_{\mathcal{X}_j}(z_j^t) = 0$  for  $j \in \mathcal{S}^t$ ,

from (52) and (53) it holds that

$$\tilde{\ell}_j(\tilde{\mathbf{x}}_j, z_j^{t+1}, \tilde{\mathbf{y}}_j) - \tilde{\ell}_j(\tilde{\mathbf{x}}_j, z_j^t, \tilde{\mathbf{y}}_j) \leq -\frac{\rho_j}{2} \|z_j^{t+1} - z_j^t\|^2 \quad (54)$$

For all  $j \notin \mathcal{S}^t$ , since  $z_j^{t+1} = z_j^t$ , the right-hand side is clearly zero. Finally, the result in (36) is obtained by summing both sides in (54) over  $j = 1, 2, \dots, K$ . ■

*Proof of Lemma 1(a):* From Lemma 2, the decrease in the Lagrangian over consecutive time slots is given by

$$\begin{aligned} & L(\{\mathbf{x}_k^{t+1}\}, \mathbf{z}^{t+1}, \{\mathbf{y}_k^{t+1}\}) - L(\{\mathbf{x}_k^t\}, \mathbf{z}^t, \{\mathbf{y}_k^t\}) \\ & \leq -\sum_{k=1}^K \frac{\rho_k - 7L_k}{2} \sum_{j \in \mathcal{N}_k} \|x_{kj}^{t+1} - x_{kj}^t\|^2 - \sum_{k \in \mathcal{S}^t} \frac{\rho_k}{2} \|z_k^{t+1} - z_k^t\|^2 \\ & + \sum_{k=1}^K \left( \frac{1}{\rho_k} + \frac{7L_k}{2\rho_k^2} \right) L_k^2(T_k+1) \sum_{d=0}^{T_k} \sum_{j \in \mathcal{N}_k} \|z_j^{t+1-d} - z_j^{t-d}\|^2 \\ & + \sum_{k=1}^K \frac{L_k T_k}{2} \sum_{d=0}^{T_k-1} \sum_{j \in \mathcal{N}_k} \|z_j^{t+1-d} - z_j^{t-d}\|^2 \quad (55) \end{aligned}$$

Given  $t_0 \geq 1$  and  $T \geq 1$ , define the inverse mapping  $\mathcal{S}_T^{-1}(k) := \{t_0 \leq t \leq t_0 + T \mid k \in \mathcal{S}^t\}$  as the set of iterations for which update (9) is applied at node  $k$  and let  $f_k \geq |\mathcal{S}_T^{-1}(k)|/T$ . Then, summing both sides of (55) over  $t = t_0, t_0 + 1, \dots, T + t_0$ ,

$$\begin{aligned} & L(\{\mathbf{x}_k^{T+t_0}\}, \mathbf{z}^{T+t_0}, \{\mathbf{y}_k^{T+t_0}\}) - L(\{\mathbf{x}_k^{t_0}\}, \mathbf{z}^{t_0}, \{\mathbf{y}_k^{t_0}\}) \leq \\ & -\sum_{t=t_0}^{T+t_0} \sum_{k=1}^K \frac{\beta_k}{2} \sum_{j \in \mathcal{N}_k} \|x_{kj}^{t+1} - x_{kj}^t\|^2 + \alpha_k \|z_k^{t+1} - z_k^t\|^2 \quad (56) \end{aligned}$$

where  $\alpha_k$  and  $\beta_k$  are defined as in (19), yielding the desired result. Note that from Assumption (A4), the term on the right-hand side of (56) is negative. ■

*Proof of Lemma 1(b):* Using the Lipschitz continuity of  $\nabla g_k(\cdot)$  and applying triangle inequality, it follows that

$$\begin{aligned} g_k(\mathbf{z}_k^{t+1}) & \leq g_k(\mathbf{x}_k^{t+1}) + \langle \nabla g_k(\mathbf{x}_k^{t+1}), \mathbf{z}_k^{t+1} - \mathbf{x}_k^{t+1} \rangle \\ & + \frac{L_k}{2} \|\mathbf{x}_k^{t+1} - \mathbf{z}_k^{t+1}\|^2 \quad (57) \end{aligned}$$

$$\begin{aligned} & = g_k(\mathbf{x}_k^{t+1}) + \langle \nabla g_k(\mathbf{x}_k^{t+1}) - \nabla g_k(\mathbf{z}_k^{t+1}), \mathbf{z}_k^{t+1} - \mathbf{x}_k^{t+1} \rangle \\ & + \langle \nabla g_k(\mathbf{z}_k^{t+1}), \mathbf{z}_k^{t+1} - \mathbf{x}_k^{t+1} \rangle + \frac{L_k}{2} \|\mathbf{x}_k^{t+1} - \mathbf{z}_k^{t+1}\|^2 \quad (58) \end{aligned}$$

$$\begin{aligned} & \leq g_k(\mathbf{x}_k^{t+1}) + \langle \nabla g_k(\mathbf{z}_k^{t+1}), \mathbf{z}_k^{t+1} - \mathbf{x}_k^{t+1} \rangle \\ & + \frac{3L_k}{2} \|\mathbf{x}_k^{t+1} - \mathbf{z}_k^{t+1}\|^2 \quad (59) \end{aligned}$$

Next, using the relationship from (37), observe that the

$$\begin{aligned} & L(\{\mathbf{x}_k^{t+1}\}, \mathbf{z}^{t+1}, \{\mathbf{y}_k^{t+1}\}) \\ & = \sum_{k=1}^K \left( g_k(\mathbf{x}_k^{t+1}) + \sum_{j \in \mathcal{N}_k} \langle \mathbf{y}_{kj}^{t+1}, x_{kj}^{t+1} - z_j^{t+1} \rangle \right. \\ & \quad \left. + \frac{\rho_k}{2} \sum_{j \in \mathcal{N}_k} \|x_{kj}^{t+1} - z_j^{t+1}\|^2 + h_k(z_k^{t+1}) \right) \quad (60) \end{aligned}$$

$$\begin{aligned} & \geq \sum_{k=1}^K \left( g_k(\mathbf{z}_k^{t+1}) + h_k(z_k^{t+1}) \right. \\ & \quad \left. + \frac{\rho_k - 3L_k}{2} \sum_{j \in \mathcal{N}_k} \|x_{kj}^{t+1} - z_j^{t+1}\|^2 \right. \\ & \quad \left. + \langle \nabla g_k(\mathbf{z}_k^{t+1}) - \nabla g_k(\mathbf{z}_k^{[t+1]}), \mathbf{z}_k^{t+1} - \mathbf{x}_k^{t+1} \rangle \right) \quad (61) \end{aligned}$$

Next, using Cauchy-Schwarz inequality on the last term, it follows that

$$\begin{aligned} & L(\{\mathbf{x}_k^{t+1}\}, \mathbf{z}^{t+1}; \{\mathbf{y}_k^{t+1}\}) \\ & \geq P + \frac{\rho_k - 3L_k}{2} \sum_{j \in \mathcal{N}_k} \left\| x_{kj}^{t+1} - z_j^{t+1} \right\|^2 \\ & \quad - \left\| \nabla g_k(\mathbf{z}_k^{t+1}) - \nabla g_k(\mathbf{z}_k^{[t+1]}) \right\| \left\| \mathbf{z}_k^{t+1} - \mathbf{x}_k^{t+1} \right\| \quad (62) \end{aligned}$$

$$\begin{aligned} & \geq P + \frac{\rho_k - 4L_k}{2} \sum_{j \in \mathcal{N}_k} \left\| x_{kj}^{t+1} - z_j^{t+1} \right\|^2 \\ & \quad - \frac{L_k}{2} \sum_{j \in \mathcal{N}_k} \left\| z_j^{t+1} - z_j^{[t+1]} \right\|^2 \quad (63) \end{aligned}$$

$$\geq P - \frac{L_k}{2} \sum_{j \in \mathcal{N}_k} \text{diam}^2(\mathcal{X}_j) > -\infty \quad (64)$$

where the last inequality follows from the fact that  $\rho_k \geq 7L_k$  [cf. (A4)], and that  $\mathcal{X}_j$  is compact [cf. (A2)]. ■

## APPENDIX B PROOF OF THEOREM 1

*Proof of Theorem 1(a):* From Lemma 1, it follows that  $L(\{\mathbf{x}_k^t\}, \mathbf{z}^t; \{\mathbf{y}_k^t\})$  converges as  $t \rightarrow \infty$ . Therefore, it holds from (20) that,

$$\lim_{t \rightarrow \infty} \left\| x_{kj}^{t+1} - x_{kj}^t \right\| \rightarrow 0 \quad j \in \mathcal{N}_k, \quad k = 1, \dots, K \quad (65a)$$

$$\lim_{t \rightarrow \infty} \left\| z_j^{t+1} - z_j^t \right\| \rightarrow 0 \quad k = 1, \dots, K \quad (65b)$$

Next, using (65b) into (38) and the form of the update in (15), we obtain

$$\lim_{t \rightarrow \infty} \left\| y_{kj}^{t+1} - y_{kj}^t \right\| \rightarrow 0 \quad j \in \mathcal{N}_k, \quad k = 1, \dots, K \quad (65c)$$

$$\lim_{t \rightarrow \infty} \left\| x_{kj}^t - z_j^t \right\| \rightarrow 0 \quad j \in \mathcal{N}_k, \quad k = 1, \dots, K. \quad (65d)$$

*Proof of Theorem 1(b):* From (65), it follows that the limit points  $\{\{x_{kj}^*\}, \{z_j^*\}, \{y_{kj}^*\}\}$  exist, and satisfy

$$x_{kj}^* = z_j^* \quad j \in \mathcal{N}_k, \quad k = 1, \dots, K \quad (66)$$

which is the primal feasibility condition for (5). Since  $z_j^{t+1} \in \mathcal{X}_j$  for all  $t$ , it should hold that  $z_j^* \in \mathcal{X}_j$  and  $x_{kj}^* = x_{lj}^*$  for all  $j, l, k \in \mathcal{N}_k$ . The first order optimality condition for  $\mathbf{x}_k$  can be obtained from (37), which implies (22a). Finally, the first order optimality condition for  $z_j$  can be obtained from (52), which implies that

$$0 \in \partial(h_j(z) + \mathbf{1}_{z \in \mathcal{X}_j})|_{z=z_j^*} - \sum_{k \in \mathcal{N}_j} y_{kj}^* - \sum_{k \in \mathcal{N}_j} \rho_k (x_{kj}^* - z_j^*).$$

The desired result follows by using (66) and noting that  $0 \in \partial \mathbf{1}_{z \in \mathcal{X}_j}|_{z=z_j^*}$ . ■

## APPENDIX C CONVERGENCE OF THE MAJORIZED ADMM

The proof of convergence of Theorem 2 follows the same general outline as that of Theorem 1. The definitions of  $\ell_k$  and  $\tilde{\ell}_j$  are the same as in (29), while the approximation of  $\ell_k(\mathbf{x}_k, \mathbf{y}_k, \mathbf{z}_k)$  at  $\mathbf{x}_k = \tilde{\mathbf{z}}$  is now defined as

$$\begin{aligned} u_k(\mathbf{x}_k, \tilde{\mathbf{z}}, \mathbf{z}_k, \mathbf{y}_k) &:= f_k(\mathbf{x}_k, \tilde{\mathbf{z}}) + \sum_{j \in \mathcal{N}_k} \langle y_{kj}, x_{kj} - z_j \rangle \\ &+ \sum_{j \in \mathcal{N}_k} \frac{\rho_k}{2} \|x_{kj} - z_j\|^2. \quad (67) \end{aligned}$$

From the update in (25), it holds that  $\nabla_{\mathbf{x}_k} u_k(\mathbf{x}_k^{t+1}, \mathbf{z}_k^{[t+1]}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t) = 0$ , and

$$[\nabla f_k(\mathbf{x}_k^{t+1}, \mathbf{z}_k^{[t+1]})]_j + y_{kj}^t + \rho_k(x_{kj} - z_j^{t+1}) = 0 \quad (68)$$

Further, from the update of  $y_{kj}^{t+1}$  and from (68),

$$y_{kj}^{t+1} = -[\nabla_{\mathbf{x}_k} f_k(\mathbf{x}_k^{t+1}, \mathbf{z}_k^{[t+1]})]_j, \quad j \in \mathcal{N}_k. \quad (69)$$

Similarly, it holds that  $y_{kj}^t = -[\nabla_{\mathbf{x}_k} f_k(\mathbf{x}_k^t, \mathbf{z}_k^{[t]})]_j$  for all  $j \in \mathcal{N}_k$ . Therefore, for each  $k \in \mathcal{K}$ , the following bound applies:

$$\begin{aligned} \sum_{j \in \mathcal{N}_k} \left\| y_{kj}^{t+1} - y_{kj}^t \right\|^2 &= \left\| \nabla f_k(\mathbf{x}_k^{t+1}, \mathbf{z}_k^{[t+1]}) - \nabla f_k(\mathbf{x}_k^t, \mathbf{z}_k^{[t]}) \right\|^2 \\ &\leq 2L_k^2 \left\| \mathbf{x}_k^{t+1} - \mathbf{x}_k^t \right\|^2 + 2L_k^2 \left\| \mathbf{z}_k^{[t+1]} - \mathbf{z}_k^{[t]} \right\|^2 \quad (70a) \\ &\leq 2L_k^2 \left\| \mathbf{x}_k^{t+1} - \mathbf{x}_k^t \right\|^2 + 2L_k^2(T_k + 1) \sum_{d=0}^{T_k} \sum_{j \in \mathcal{N}_k} \left\| z_j^{t+1-d} - z_j^{t-d} \right\|^2 \quad (70b) \end{aligned}$$

where (70a) follows from Assumption, and (70b) follows similarly as in (38c).

As in Appendix A, the Lagrangian is split into three summands [cf. Lemma 2], each of which must be separately bounded. The bound on the first summand follows directly from (70).

$$\begin{aligned} & L(\{\mathbf{x}_k^{t+1}, \mathbf{z}^{t+1}, \{\mathbf{y}_k^{t+1}\}\}) - L(\{\mathbf{x}_k^{t+1}\}, \mathbf{z}^{t+1}, \{\mathbf{y}_k^t\}) \\ &= \sum_{k=1}^K \frac{1}{\rho_k} \sum_{j \in \mathcal{N}_k} \left\| y_{kj}^{t+1} - y_{kj}^t \right\|^2 \quad (71) \\ &\leq \sum_{k=1}^K \frac{2L_k^2}{\rho_k} \sum_{j \in \mathcal{N}_k} \left\| x_{kj}^{t+1} - x_{kj}^t \right\|^2 \\ &\quad + \sum_{k=1}^K \frac{2L_k^2(T_k + 1)}{\rho_k} \sum_{d=0}^{T_k} \sum_{j \in \mathcal{N}_k} \left\| z_j^{t+1-d} - z_j^{t-d} \right\|^2. \quad (72) \end{aligned}$$

For the second summand, some relevant inequalities are first stated. From the definition of the majorizing function, it holds that for all  $\mathbf{x}_k$ ,  $\ell_k(\mathbf{x}_k, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t) \leq u_k(\mathbf{x}_k, \mathbf{z}_k^{t+1}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t)$ . Since  $f_k(\mathbf{x}_k, \mathbf{z}_k^{t+1})$  is differentiable and convex with respect to  $\mathbf{x}_k$ , it follows that  $u_k(\mathbf{x}_k, \tilde{\mathbf{z}}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t)$  is differentiable and strongly convex, implying that

$$\begin{aligned} & u_k(\mathbf{x}_k^{t+1}, \tilde{\mathbf{z}}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t) - u_k(\mathbf{x}_k^t, \tilde{\mathbf{z}}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t) \\ &\leq \langle \nabla_{\mathbf{x}_k} u_k(\mathbf{x}_k^{t+1}, \tilde{\mathbf{z}}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t), \mathbf{x}_k^{t+1} - \mathbf{x}_k^t \rangle \\ &\quad - \frac{\rho_k}{2} \left\| \mathbf{x}_k^{t+1} - \mathbf{x}_k^t \right\|^2 \quad (73) \end{aligned}$$

for all  $\tilde{\mathbf{z}}$ . In particular for  $\tilde{\mathbf{z}} = \mathbf{z}_k^{[t+1]}$ , the gradient on the right is zero from (68), so that  $u_k(\mathbf{x}_k^{t+1}, \mathbf{z}_k^{[t+1]}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t) - u_k(\mathbf{x}_k^t, \mathbf{z}_k^{[t+1]}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t) \leq -\frac{\rho_k}{2} \sum_{j \in \mathcal{N}_k} \left\| x_{kj}^{t+1} - x_{kj}^t \right\|^2$ .

Note also that  $u_k(\mathbf{x}_k^{t+1}, \mathbf{z}_k^{[t+1]}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t) - \nabla_{\mathbf{x}_k} u_k(\mathbf{x}_k^{t+1}, \mathbf{z}_k^{[t+1]}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t) = \nabla_{\mathbf{x}_k} f_k(\mathbf{x}_k^{t+1}, \mathbf{z}_k^{[t+1]}) - \nabla_{\mathbf{x}_k} f_k(\mathbf{x}_k^{t+1}, \mathbf{z}_k^{[t+1]})$ . Next, specializing (73) for  $\tilde{\mathbf{z}} = \mathbf{z}_k^{[t+1]}$ , the following series of inequalities are obtained

$$\begin{aligned} & u_k(\mathbf{x}_k^{t+1}, \mathbf{z}_k^{[t+1]}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t) - u_k(\mathbf{x}_k^t, \mathbf{z}_k^{[t+1]}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t) \\ &\leq \langle \nabla_{\mathbf{x}_k} u_k(\mathbf{x}_k^{t+1}, \mathbf{z}_k^{[t+1]}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t), \mathbf{x}_k^{t+1} - \mathbf{x}_k^t \rangle \\ &\quad - \frac{\rho_k}{2} \left\| \mathbf{x}_k^{t+1} - \mathbf{x}_k^t \right\|^2 \quad (74a) \end{aligned}$$



$$\begin{aligned} &\leq \frac{L_k T_k}{2} \sum_{d=0}^{T_k-1} \sum_{j \in \mathcal{N}_k} \|z_j^{t+1-d} - z_j^{t-d}\|^2 \\ &\quad - \frac{\rho_k - L_k}{2} \sum_{j \in \mathcal{N}_k} \|x_{kj}^{t+1} - x_{kj}^t\|^2 \end{aligned} \quad (74b)$$

Utilizing (68) and following steps (43a)-(43d), (74b) is obtained, which follow similarly as in (43).

The Lipschitz continuity of  $\nabla g_k(\cdot)$  and  $f_k(\cdot, \mathbf{z}_k^{t+1})$  also imply the following upper bounds

$$\begin{aligned} g_k(\mathbf{z}_k^{t+1}) &\leq g_k(\mathbf{x}_k^t) + \langle \nabla g_k(\mathbf{x}_k^t), \mathbf{z}_k^{t+1} - \mathbf{x}_k^t \rangle \\ &\quad + \frac{L_k}{2} \sum_{j \in \mathcal{N}_k} \|x_{kj}^t - z_j^{t+1}\|^2 \end{aligned} \quad (75)$$

$$\begin{aligned} f_k(\mathbf{x}_k^t, \mathbf{z}_k^{t+1}) &\leq f_k(\mathbf{z}_k^{t+1}, \mathbf{z}_k^{t+1}) + \frac{L_k}{2} \sum_{j \in \mathcal{N}_k} \|x_{kj}^t - z_j^{t+1}\|^2 \\ &\quad + \langle \nabla_{\mathbf{x}_k} f_k(\mathbf{z}_k^{t+1}, \mathbf{z}_k^{t+1}), \mathbf{x}_k^t - \mathbf{z}_k^{t+1} \rangle \\ &\leq g_k(\mathbf{x}_k^t) + \langle \nabla g_k(\mathbf{z}_k^{t+1}) - \nabla g_k(\mathbf{x}_k^t), \mathbf{x}_k^t - \mathbf{z}_k^{t+1} \rangle \end{aligned} \quad (76)$$

$$+ L_k \sum_{j \in \mathcal{N}_k} \|x_{kj}^t - z_j^{t+1}\|^2 \quad (77)$$

Therefore, from (77), it follows that

$$\begin{aligned} u_k(\mathbf{x}_k^t, \mathbf{z}_k^{t+1}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t) - \ell_k(\mathbf{x}_k^t, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t) \\ = f_k(\mathbf{x}_k^t, \mathbf{z}_k^{t+1}) - g_k(\mathbf{x}_k^t) \end{aligned} \quad (78a)$$

$$\begin{aligned} &\leq \langle \nabla g_k(\mathbf{z}_k^{t+1}) - \nabla g_k(\mathbf{x}_k^t), \mathbf{x}_k^t - \mathbf{z}_k^{t+1} \rangle \\ &\quad + L_k \sum_{j \in \mathcal{N}_k} \|x_{kj}^t - z_j^{t+1}\|^2 \end{aligned} \quad (78b)$$

$$\leq 2L_k \sum_{j \in \mathcal{N}_k} \|x_{kj}^t - z_j^{t+1}\|^2 \quad (78c)$$

$$\leq 4L_k \sum_{j \in \mathcal{N}_k} \left( \|x_{kj}^t - x_{kj}^{t+1}\|^2 + \|x_{kj}^{t+1} - z_j^{t+1}\|^2 \right) \quad (78d)$$

where (78c) follows from (A4) and (78d) from the use of triangle inequality. Finally, it holds that

$$\ell_k(\mathbf{x}_k^{t+1}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t) - \ell_k(\mathbf{x}_k^t, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t) \quad (79)$$

$$\begin{aligned} &\leq u_k(\mathbf{x}_k^{t+1}, \mathbf{z}_k^{t+1}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t) - u_k(\mathbf{x}_k^t, \mathbf{z}_k^{t+1}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t) \\ &\quad + u_k(\mathbf{x}_k^t, \mathbf{z}_k^{t+1}, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t) - \ell_k(\mathbf{x}_k^t, \mathbf{z}_k^{t+1}, \mathbf{y}_k^t) \end{aligned} \quad (80)$$

$$\leq -\frac{\rho_k - 9L_k}{2} \sum_{j \in \mathcal{N}_k} \|x_{kj}^{t+1} - x_{kj}^t\|^2 + \frac{4L_k}{\rho_k^2} \sum_{j \in \mathcal{N}_k} \|y_{kj}^{t+1} - y_{kj}^t\|^2$$

$$+ \frac{L_k T_k}{2} \sum_{d=0}^{T_k-1} \sum_{j \in \mathcal{N}_k} \|z_j^{t+1-d} - z_j^{t-d}\|^2 \quad (81)$$

$$= -\left( \frac{\rho_k - 9L_k}{2} - \frac{8L_k^3}{\rho_k^2} \right) \sum_{j \in \mathcal{N}_k} \|x_{kj}^{t+1} - x_{kj}^t\|^2$$

$$+ \left( \frac{8L_k^3(T_k + 1)}{\rho_k^2} + \frac{L_k T_k}{2} \right) \sum_{d=0}^{T_k-1} \sum_{j \in \mathcal{N}_k} \|z_j^{t+1-d} - z_j^{t-d}\|^2 \quad (82)$$

where (81) follows from using the update for  $y_{kj}^{t+1}$  and (82) from (70). Finally, summing (82) over  $k = 1, 2, \dots, K$ , the required bound is obtained. Since the expression for the third summand in (36) remains the same, the decrease in the Lagrangian over consecutive time slots  $t = t_0 + 1, \dots, t_0 + T$ , is given by

$$L(\{\mathbf{x}_k^{T+t_0}\}, \mathbf{z}^{T+t_0}, \{\mathbf{y}_k^{T+t_0}\}) - L(\{\mathbf{x}_k^{t_0}\}, \mathbf{z}^{t_0}, \{\mathbf{y}_k^{t_0}\})$$

$$\leq - \sum_{t=t_0}^{T+t_0} \sum_{k=1}^K \beta_k \sum_{j \in \mathcal{N}_k} \|x_{kj}^{i+1} - x_{kj}^i\|^2 + \alpha_k \|z_k^{i+1} - z_k^i\|^2 \quad (83)$$

where,  $\alpha_k$  and  $\beta_k$  are given in (27) yielding the desired result. Note that from Assumption (A5), the term on the right-hand side of (56) is negative. Next, the boundedness of the Lagrangian follows as shown in Appendix A. Finally, it is possible to apply Theorem 1 to this case, yielding the desired result.

## APPENDIX D

### LIPSCHITZ CONTINUITY OF OBJECTIVE FUNCTION IN (4)

Recall that  $g_k(\{\mathbf{x}_j\}_{j \in \mathcal{N}_k}) = \sum_j w_{kj} (\delta_{kj} - d_{kj}(\mathbf{x}_k, \mathbf{x}_j))^2$ , where the modified definition of  $d_{kj}(\mathbf{x}_k, \mathbf{x}_j) = \sqrt{\|\mathbf{x}_k - \mathbf{x}_j\| + \epsilon}$  is utilized and for a pair of node  $(k, j)$  weight  $w_{kj} = 1$  for  $j \in \mathcal{N}_k$  otherwise 0. The gradient of  $g_k$  is given by  $\nabla_{\mathbf{x}_l} g_k(\{\mathbf{x}_j\}_{j \in \mathcal{N}_k}) =$

$$\begin{cases} 2w_{kj}(\delta_{kj} - d_{kj}(\mathbf{x}_k, \mathbf{x}_j)) \frac{(\mathbf{x}_k - \mathbf{x}_j)}{d_{kj}(\mathbf{x}_k, \mathbf{x}_j)} & l \neq k \\ -\sum_j 2w_{kj}(\delta_{kj} - d_{kj}(\mathbf{x}_k, \mathbf{x}_j)) \frac{(\mathbf{x}_k - \mathbf{x}_j)}{d_{kj}(\mathbf{x}_k, \mathbf{x}_j)} & l = k \end{cases} \quad (84)$$

In order to show that the expression in (84) is Lipschitz, it suffices to prove that each element of the Hessian is bounded. Denoting the coordinate vector of the  $i$ -th node as  $\mathbf{x}_i := [x_i^1 \ x_i^2]$  and defining the  $2 \times 2$  matrix  $[C(\mathbf{x}_k, \mathbf{x}_j)]_{pq} := 2w_{kj}(1 - \delta_{kj}((x_k^p - x_j^p)(x_k^q - x_j^q) + \epsilon)d_{kj}^{-3}(\mathbf{x}_k, \mathbf{x}_j))$  for  $1 \leq p, q \leq 2$ , it can be seen that the  $(\ell, m)$ -th block of the Hessian matrix is given by

$$\nabla_{\mathbf{x}_l, \mathbf{x}_m}^2 g_k(\{\mathbf{x}_j\}_{j \in \mathcal{N}_k}) = \begin{cases} C(\mathbf{x}_k, \mathbf{x}_m) & m = l, l \neq k \\ -C(\mathbf{x}_k, \mathbf{x}_m), & m \neq l, m = k \\ \sum_{j \in \mathcal{N}_k} C(\mathbf{x}_k, \mathbf{x}_j) & m = l, l = k \\ -C(\mathbf{x}_k, \mathbf{x}_l) & m \neq l, l = k \\ 0 & m \neq l, l \neq k \end{cases}$$

For the present localization example, we have  $\max_{k,j} \{\delta_{kj}\} \leq 1$ ,  $\max_k |\mathcal{N}_k| \leq N$ ,  $w_{kj} \leq 1$ ,  $\|\mathbf{x}_k\| \leq 1 \ \forall \ k = 1, \dots, K$ , and  $\|\frac{1}{d_{kj}(\mathbf{x}_k, \mathbf{x}_j)}\| \leq \frac{1}{\sqrt{\epsilon}}$ . The application of triangle inequality therefore implies that  $\|C(\mathbf{x}_k, \mathbf{x}_m)\| \leq 8(2/\sqrt{\epsilon} + 1)$  for all  $1 \leq k, m \leq N$ . Therefore it follows that  $\nabla_{\mathbf{x}_l, \mathbf{x}_m}^2 g_k(\{\mathbf{x}_j\}_{j \in \mathcal{N}_k})$  is bounded, as claimed.

## REFERENCES

- [1] T. W. Martin and K.-C. Chang, "A distributed data fusion approach for mobile ad hoc networks," in *Proc. of the Int. Conf. on Info. Fusion*, 2005, vol. 2. Philadelphia, PA, USA: IEEE, 2005, pp. 8–pp.
- [2] A. Speranzon, C. Fischione, and K. H. Johansson, "Distributed and collaborative estimation over wireless sensor networks," in *Proc. of IEEE CDC*. San Diego, CA, USA: IEEE, 2006, pp. 1025–1030.
- [3] A. Simonetto and G. Leus, "Distributed maximum likelihood sensor network localization," *IEEE Trans. Signal Process.*, vol. 62, no. 6, pp. 1424–1437, 2014.
- [4] J. A. Costa, N. Patwari, and A. O. Hero III, "Distributed weighted-multidimensional scaling for node localization in sensor networks," *ACM Trans. Sensor Netw.*, vol. 2, no. 1, pp. 39–64, 2006.
- [5] G. Mateos and K. Rajawat, "Dynamic network cartography: Advances in network health monitoring," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 129–143, 2013.
- [6] J. Cortés, "Distributed kriged kalman filter for spatial estimation," *IEEE Trans. Autom. Control*, vol. 54, no. 12, pp. 2816–2827, 2009.
- [7] P. A. Forero, A. Cano, and G. B. Giannakis, "Consensus-based k-means algorithm for distributed learning using wireless sensor networks," in *Proc. of the Workshop on Sensors, Signal and Info. Process.*, Sedona, AZ, 2008, pp. 11–14.

- [8] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trend. Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.
- [9] J. C. Duchi, A. Agarwal, and M. J. Wainwright, "Dual averaging for distributed optimization: convergence analysis and network scaling," *IEEE Trans. Autom. Control*, vol. 57, no. 3, pp. 592–606, 2012.
- [10] N. Parikh and S. Boyd, "Proximal algorithms," *Found. Trend. Optim.*, vol. 1, no. 3, pp. 123–231, 2013.
- [11] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Trans. Autom. Control*, vol. 54, no. 1, pp. 48–61, 2009.
- [12] I. Lobel and A. Ozdaglar, "Distributed subgradient methods for convex optimization over random networks," *IEEE Trans. Autom. Control*, vol. 56, no. 6, pp. 1291–1306, 2011.
- [13] D. P. Palomar and Y. C. Eldar, *Convex optimization in signal processing and communications*. Cambridge university press, 2010.
- [14] H. Terelius, U. Topcu, and R. Murray, "Decentralized multi-agent optimization via dual decomposition," in *Proc. of the IFAC World Congress, 2011*, Milano, Italy, 2011, pp. 11 245–11 251.
- [15] E. Wei and A. Ozdaglar, "On the  $o(1/k)$  convergence of asynchronous distributed alternating direction method of multipliers," *arXiv preprint*, 2013. [Online]. Available: <http://arxiv.org/pdf/1307.8254.pdf>
- [16] M. Hong, "A distributed, asynchronous and incremental algorithm for nonconvex optimization: An ADMM based approach," *arXiv preprint*, 2014. [Online]. Available: <http://arxiv.org/pdf/1412.6058.pdf>
- [17] M. Hong, Z.-Q. Luo, and M. Razaviyayn, "Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems," *arXiv preprint*, 2014. [Online]. Available: <http://arxiv.org/pdf/1410.1390.pdf>
- [18] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and distributed computation: numerical methods*. Prentice hall Englewood Cliffs, NJ, 1989, vol. 23.
- [19] J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Trans. Autom. Control*, vol. 31, no. 9, pp. 803–12, 1986.
- [20] M. Li, D. G. Andersen, and A. Smola, "Distributed delayed proximal gradient methods," in *Proc. of NIPS Workshop on Optimization for Machine Learning*, Lake Tahoe, Nevada, USA., 2013.
- [21] A. Agarwal and J. C. Duchi, "Distributed delayed stochastic optimization," in *Proc. of the NIPS*, Granada Spain, 2011, pp. 873–881.
- [22] J. Liu, S. J. Wright, C. Ré, V. Bittorf, and S. Sridhar, "An asynchronous parallel stochastic coordinate descent algorithm," *arXiv preprint*, 2013. [Online]. Available: <http://arxiv.org/pdf/1311.1873.pdf>
- [23] G. Scutari, D. P. Palomar, and S. Barbarossa, "Asynchronous iterative water-filling for gaussian frequency-selective interference channels," *IEEE Trans. Inf. Theory*, vol. 54, no. 7, pp. 2868–2878, 2008.
- [24] A. Nedic, D. P. Bertsekas, and V. S. Borkar, "Distributed asynchronous incremental subgradient methods," *Studies in Computational Mathematics*, vol. 8, pp. 381–407, 2001.
- [25] R. Zhang and J. Kwok, "Asynchronous distributed ADMM for consensus optimization," in *Proc. of the ICML*, Beijing, China, 2014, pp. 1701–09.
- [26] K. Srivastava and A. Nedić, "Distributed asynchronous constrained stochastic optimization," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 4, pp. 772–790, 2011.
- [27] X. Zhao and A. H. Sayed, "Asynchronous adaptation and learning over networks: part i: Modeling and stability analysis," *IEEE Trans. Signal Process.*, vol. 63, no. 4, pp. 811–826, 2015.
- [28] T.-H. Chang, "A proximal dual consensus ADMM method for multi-agent constrained optimization," *arXiv preprint*, 2014. [Online]. Available: <http://arxiv.org/pdf/1409.3307.pdf>
- [29] M. Chiang, "Nonconvex optimization of communication systems," *Advances in Mechanics and Mathematics*, vol. 3, pp. 137–196, 2008.
- [30] N. Cressie, *Statistics for spatial data*. John Wiley & Sons, 2015.
- [31] V. Kekatos and G. Giannakis, "Distributed robust power system state estimation," *IEEE Trans. Power Systems*, vol. 28, no. 2, pp. 1617–1626, 2013.
- [32] D. Aleksandar and B. Zhang, "Distributed estimation and detection for sensor networks using hidden markov random field models," *IEEE Trans. Signal Process.*, vol. 54, no. 8, pp. 3200–3215, 2006.
- [33] K. Rajawat, A. Cano, and G. B. Giannakis, "Network-compressive coding for wireless sensors with correlated data," *IEEE Trans. Wireless Comm.*, vol. 11, no. 12, pp. 4264–4274, 2012.
- [34] R. W. Ouyang, A. K.-S. Wong, and C.-T. Lea, "Received signal strength-based wireless localization via semidefinite programming: Noncooperative and cooperative schemes," *IEEE Trans. Veh. Technol.*, vol. 59, no. 3, pp. 1307–1318, 2010.
- [35] X. Lian, Y. Huang, Y. Li, and J. Liu, "Asynchronous parallel stochastic gradient for nonconvex optimization," *arXiv preprint*, 2015. [Online]. Available: <http://arxiv.org/pdf/1506.08272.pdf>
- [36] S. Magnússon, P. C. Weeraddana, M. G. Rabbat, and C. Fischione, "On the convergence of alternating direction lagrangian methods for nonconvex structured optimization problems," *arXiv preprint*, 2014. [Online]. Available: <http://arxiv.org/pdf/1409.8033.pdf>
- [37] I. Borg and P. J. Groenen, *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.
- [38] F. Iutzeler, P. Bianchi, P. Ciblat, and W. Hachem, "Asynchronous distributed optimization using a randomized alternating direction method of multipliers," in *Proc. of IEEE CDC*, Firenze, Italy, 2013, pp. 3671–76.
- [39] A. Nedić, "Asynchronous broadcast-based convex optimization over a network," *IEEE Trans. Autom. Control*, vol. 56, no. 6, pp. 1337–1351, 2011.
- [40] S. Barbarossa, S. Sardellitti, and P. Di Lorenzo, "Distributed detection and estimation in wireless sensor networks," *arXiv preprint arXiv:1307.1448*, 2013.